MULTIMODAL FAITHSCORE: Evaluating Hallucinations in Large Vision-Language Models

Anonymous ACL submission

Abstract

We introduce MULTIMODAL FAITHSCORE (**Faithfulness** in Atomic Fact **Score**), a fine-grained evaluation metric that measures the faithfulness of generated responses to an input image and corresponding *open-ended questions*. MULTIMODAL FAITHSCORE first identifies sub-sentences that should be verified, then extracts nuanced elements from identified sub-sentences, and finally conducts consistency verification between elements and the input image. Meta-evaluation demonstrates that our reference-free metric highly correlates with human judgments of faithfulness. We measure hallucinations in state-of-the-art LLVMs with MULTIMODAL FAITHSCORE. Based on both automatic evaluation and human judgments, we show that current systems mostly face challenges such as unfaithful generations that are not grounded to the image, which leaves room for improvements in performance. Our codes and data are publicly available at https://577279815.wixsite.com/multimodalfathscore.

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., 2023b; 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). Furthermore, LVLMs have shown strong performance on various multimodal tasks, such as VQA (Antol et al., 2015), Image Captioning (Lin et al., 2014), and Multimodal Conversation (Liu et al., 2023b).

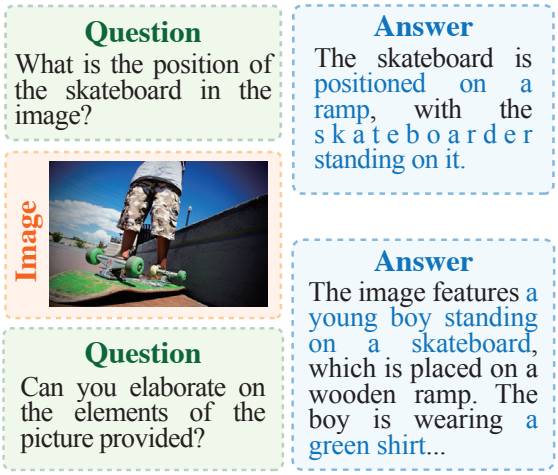

Figure 1: Two testing examples for LVLMs. Answers are generated by the LLaVA model. The part highlighted in blue denotes hallucinations in the answers.

Unfortunately, the problem of hallucination in LLMs is pervasive and poses a significant challenge, often leading these models to generate misleading or fabricated information (Ji et al., 2023b). Similar to LLMs, LVLMs continue to grapple with the issue of hallucination, where they produce inaccurate information that is incongruent with the provided visual input. As shown in Figure 1, one of the commonly used LVLMs, LLaVA (Liu et al., 2023b), generates answers with several inaccurate descriptions (*e.g., stateboarder standing on it*, *a young boy standing on a sketeboard* and *a green shirt*), which is not faithful towards the input image. The phenomenon of hallucination in models introduces potential hazards that could result in significant consequences such as misinformation and safety concerns, thus degrading the model's usability inevitably (MacLeod et al., 2017). Hence, it is imperative that these issues are thoroughly measured and addressed (Ji et al., 2023b).

Despite the phenomenon of hallucinations in LVLMs should not be underestimated, there has been limited exploration that measured the degree

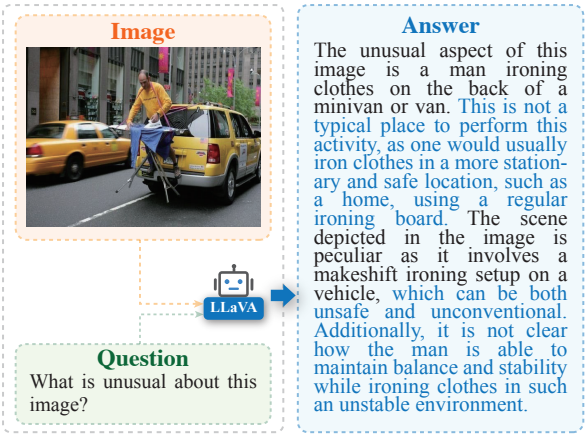

Figure 2: A testing sample of complex reasoning question for LVLMs. The blue words denotes commonsense reasoning-related content in the answer.

of hallucinations in LVLMs. Li et al. (2023b) was among the first to propose an evaluation method for the hallucinations of LVLMs. In addition, Gunjal et al. (2023) annotated a multi-modal hallucination detection dataset for detailed image description evaluation. Moreover, Lovenia et al. (2023) devised Negative Object Presence Evaluation (NOPE) to quantitatively evaluate object hallucination through VQA. Although prior work has achieved success, none of them explored evaluating hallucination of the complex and free-form response to the open-ended question (OpenAI, 2023).

Evaluating hallucinations present in free-form responses is especially challenging for two primary reasons: (1) **Free-form answers contain hybrid of description and analysis.** Unlike image captioning, answering complex questions in a free-form manner does not simply generate a descriptive content of the given image. It also contains a certain degree of analytic content such as commonsense reasoning. As depicted in Figure 2, certain sub-sentences (e.g., those highlighted in blue) do not require verification with the image input due to their analytical nature. More precisely, they encompass commonsense reasoning that extends beyond mere visual inputs, rather than solely offering a direct description of the visual modality. Thus, pinpointing the descriptive content within the responses generated by LVLMs poses a major challenge. (2) **Model outputs are prone to contain multiplicity of hallucinations.** Current methodologies offer a constricted view on evaluating hallucinations, primarily concentrating on coarse-grained object existences, while neglecting other fine-grained elements, such as counts, colors, and the spatial interrelations between objects, which also form a significant portion of hallucinations (Gunjal et al., 2023). Consequently, devising a method to holistically evaluate fine-grained hallucinations emerges as another substantial challenge.

To address the aforementioned challenges, we propose the MULTIMODAL FAITHSCOREmetric. This metric comprises three primary components: Descriptive Sub-sentence Identification, Atomic Fact Generation, and Fact Verification, as illustrated in Figure 3. The first component is tasked with discerning descriptive sub-sentences within the composite content of the generated answer. The second component deconstructs this descriptive content into fine-grained elements (*i.e.,* atomic facts). These facts cover a variety of types, such as color and count. The final component emphasizes verifying the consistency between the visual modality information and the devised atomic facts via the Visual Entailment Model (VEM) (Xie et al., 2019). Based on the proposed metric, we evaluated several advanced LVLMs. From the result, we conclude that current LVLMs still face challenges of unfaithful generations that are not grounded in the image, which leaves room for improvement.

In summary, our contributions are as follows: (1) We introduce MULTIMODAL FAITHSCORE, a metric tailored to assess various types of hallucinations in LLVMs free-form answers to open-ended questions, which is neglected by current methods; (2) To the best of our knowledge, this work is the first study that systematically evaluates the fine-grained hallucinations of free-form answers to open-ended questions in existing LVLMs; (3) In our quest to understand the hallucinations manifested by LVLMs, we embarked on comprehensive experiments with six publicly available models across diverse datasets. Our findings underscore that addressing hallucination remains a pressing challenge for LVLMs.

2 Related Work

Large Vision-Language Model Motivated by the success of the pertaining technique in LLMs and Vision Foundation Models (VFMs), the multi-modal committee has recently shifted the research attention to LVLMs. Contemporary advanced LVMs predominantly feature three core components: a text encoder, an image encoder, and a cross-modal alignment module. 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., 2023a). For instance, Liu et al. (2023b) 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.

Vision-language Model Hallucinations and Evaluations Though hallucination phenomenons and mitigation methods have been extensively studied in the text generation literature (Ji et al., 2023a; Min et al., 2023a), it is much less investigated in vision-language models (Dai et al., 2023; Liu et al., 2023c). Current work mainly focuses on the constraint setting such as image captioning. For example, Rohrbach et al. (2018) propose caption hallucination assessment 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, 2005a) and CIDEr (Vedantam et al., 2015) do not capture this. Li et al. (2023a) extends CHAIR and proposes "POPE", 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.

Different from all above, we are the first to propose a general metric for evaluating the responses of *open-ended* visual question answering setting, where answers are of free-form and lengthy (passages).

3 MULTIMODAL FAITHSCORE

In this section, we begin by clearly defining the research problem, followed by a detailed explanation of the devised MULTIMODAL FAITHSCOREmetric framework.

3.1 Task and Settings

Suppose we have an image I and a question Q corresponding to the image and then feed them into a large vision-language model denoted as $\mathcal{M}$, and obtain the generated answer A_i. Our objective is to design a scoring function $\mathcal{F}$ that yields a faithfulness score based on the generated answer A, the input question Q, and the input image I, defined as $f = \mathcal{F}(A, Q, I)$. f is a real value ranging between 0 and 1. Notably, the devised evaluation method doesn't need a ground truth answer.

In order to assess the faithfulness of the generated answers by Large Vision-Language models, we introduce a novel metric called MULTIMODAL FAITHSCOREto implement the scoring function $\mathcal{F}$. The MULTIMODAL FAITHSCOREmetric comprises three key components: descriptive sub-sentence identification, atomic fact generation, and fact verification, as depicted in Figure 3. We introduce Recognizer, Dcomposer, and Verifier, to fulfill these components, respectively.

Descriptive Sub-sentence Identification. Unlike LLMs, faithfulness in the context of vision-language models refers to the consistency between the input visual modality content and the generated answer. Notably, we should focus on the content that is an objective description of the input image. Therefore, our first step is to identify the descriptive sub-sentences within the answer by a recognizer, to obtain a more precise and fine-grained understanding of the hallucination.

Based on the actual answers generated by LVLMs, we have observed that humans are capable of distinguishing descriptive sub-sentences from other sub-sentences (referred to as analytical sub-sentences) by analyzing the content of these generated answers. However, it's important to note that manually identifying descriptive sub-sentences within the answers is a resource-intensive process, requiring both time and labor for human annotations. As a practical solution, we turn to the LLM

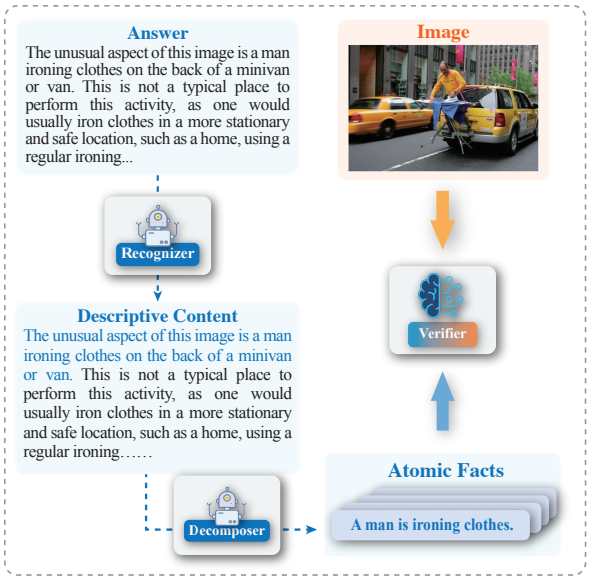

Figure 3: An overview of MULTIMODAL FAITHSCORE, which mainly consists of three components: Descriptive Sub-sentence Identification, Atomic Fact Generation, and Fact Verification. The blue words denote recognized descriptive content.

to implement the recognizer, which has demonstrated remarkable text analysis capabilities across a wide range of natural language processing tasks.

To be specific, our approach focuses on crafting a prompt that encompasses task instructions and in-context learning examples. This designed prompt is subsequently input into LLMs, leading to the generation of annotated results, defined by the equation:

$$\hat{A} = LLM(A, P) \tag{1}$$

where $\hat{A}$ signifies the generated result, from which we can extract all descriptive sub-sentences, denoted as A'. For a more comprehensive understanding of the specific prompt P utilized in this process, please refer to the Appendix.

Atomic Fact Generation. The descriptive sub-sentences denoted as A' encompass multiple pieces of information (*i.e.,* atomic facts), each of which may have varying degrees of faithfulness. Therefore, we need a decomposer to break the sub-sentences into atomic facts to get a fine-grained evaluation. In order to assess the faithfulness of answers generated by LVLMs at a fine-grained level, prior research (Li et al., 2023b) has employed the Caption Hallucination Assessment with Image Relevance (CHAIR) metric (Rohrbach et al., 2018) to evaluate hallucination in image captioning tasks. However, it is worth noting that the CHAIR metric primarily focuses on object-level hallucination, neglecting aspects such as object attributes and relationships between objects. Additionally, it relies on the availability of additional labels for objects and encompasses a limited variety of object types.

Hence, to address these limitations, we introduce a novel mechanism that can break down the original sentence into atomic facts. In particular, we break sentences down into atomic facts. We divide the atomic facts into five distinct categories: entity, count, color, relation, and other attributes, which can cover all the content of the answer. In this paper, we define atomic fact as the information belonging to one category in the above five categories. Importantly, the atomic fact is a minimal unit of information, which can ensure verification of each element in the answer and avoid being disturbed by other information. For example, for the category entity, the atomic fact can't contain more than two entities. To facilitate understanding, you can refer to the prompt of atomic fact generation in the Appendix and get some examples.

This approach allows us to identify and assess hallucinations in terms of different categories and atomic levels. To achieve this, similar to the process of identifying descriptive sub-sentences, we also rely on LLMs for the generation of atomic facts. More precisely, we annotate a set of K examples for demonstrations and prompt LLMs for sentence decomposition as follows:

$$E_i = LLM(A', P'), i \in [1, C] \tag{2}$$

where $E_i = \{e_i^1, \cdots, e_i^{n_i}\}$ represents all (*i.e.,* n_i) atomic facts pertaining to the i-th category, and C stands for the total number of categories. It is important to note that the set E_i may occasionally be an empty set. Further details regarding the specific prompt P' utilized in this process can be found in the Appendix.

Fact Verification. To calculate the MULTIMODAL FAITHSCOREscore for the LVLM answer, we first compute the score for each fact and then aggregate them to derive the overall score using the following formula:

$$\hat{f} = \frac{\sum_{i=1}^{C} \sum_{j=1}^{n_j} w_i^j \cdot s(e_i^j, I)}{\sum_{i=1}^{C} \sum_{j=1}^{n_j} 1}, \tag{3}$$

where $\hat{f}$ represents the overall faithfulness score of the answer A. The function $s(e_i^j, I)$ refers to

Metric	Spearman's ρ %	Kendall's τ %
BLEU-1	4.6	-10.3
BLEU-2	2.0	-9.2
BLEU-3	-2.6	-12.6
BLEU-4	-4.6	-13.5
ROUGE-1	22.1	14.6
ROUGE-2	9.3	-1.6
ROUGE-L	14.1	10.3
METEOR	0.6	-12.5
CHAIR	-26.8	-38.5
CLIP-Score	-6.5	-1.5
Ours	34.8	28.5

Table 1: Correlation between each evaluation metric and human judgment on LVLM hallucinations, measured by Spearman's ρ and Kendall's τ.

the verification function (*i.e.,* Verifier), which determines whether e_i^j can be supported by the input image I. To implement this function, we resort to the Visual Entailment model. When the Visual Entailment model outputs 1, indicating that the image I semantically entails the text e_i^j, and 0 otherwise. The parameter w_i^j is a weighted factor that can be used to assign different weights to different atomic facts for various tasks. In this paper, we set all the weights to 1, following the setting of the existing work (Min et al., 2023b; Krishna et al., 2023).

In addition, we further introduce a sentence-level MULTIMODAL FAITHSCOREscore as follows,

$$\hat{f}_s = \frac{C_h}{C}, \tag{4}$$

where C is the total number of descriptive sub-sentences in the answer and C_h is the total number of descriptive sub-sentences with hallucination in the answer.

4 Meta-evaluate MULTIMODAL FAITHSCORE for Automatic Evaluation

To verify that our automatic evaluation correlates with human judgment, we conduct a human evaluation in terms of hallucination. We select the testing dataset from LLaVA for human evaluation. This testing set is a visual instruction dataset comprising three distinct sample types: detailed description, conversation, and complex question. For each of these sample types, this dataset included 30 samples. We select LLaVA (Liu et al., 2023b) and mPLUG-Owl (Ye et al., 2023) models for hallucination evaluation.

4.1 Annotation Process

For each testing sample, we meticulously crafted an annotation process to assign the faithfulness score to the text via the subsequent steps.

Step 1: Sub-sentence Identification. Annotators first should 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 "description" category; otherwise, it's categorized as "analytics". For the "analytics" sub-sentence, annotators should skip the following steps. Otherwise, they should follow the next steps.

Step 2: Atomic Fact Revision. In this step, the human annotator should decompose the descriptive sub-sentence 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 meticulously examine each atomic fact to ensure its fidelity to the given sub-sentence. Atomic facts that are either redundant or non-atomic facts are promptly 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 intended entity or concept by revision atomic facts. Additionally, any missing atomic facts that should have been included in the sub-sentence are added.

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 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 correlation score, we employed the Likert Scale (Likert, 1932) to gauge the faithfulness of LVLM. This approach transforms human evaluations into a tangible scale, ranging from 1 (being the poorest) to 5 (being the best). The details are given in the Appendix B.

We have 3 employers for annotation and every person annotated 180 testing samples. We recruit

Model	Conversation	Detailed Description	Complex Question	Overall
MiniGPT-4	0.7122	0.7006	0.7019	0.7049
LLaVA	0.7403	0.7164	0.7163	0.7243
LLaVA-1.5	0.7419	0.7378	0.7358	0.7385
InstructBLIP	0.7650	0.7399	0.7439	0.7496
Multimodal-GPT	0.6458	0.6314	0.6387	0.6386
mPLUG-Owl	0.7868	0.7696	0.7778	0.7781

Table 2: Evaluation results of different LVLMs on the LLaVA-1k dataset.

Model	Performance
MiniGPT-4	0.6482
LLaVA	0.6760
LLaVA-1.5	0.7689
InstructBLIP	0.8316
Multimodal-GPT	0.6710
mPLUG-Owl	0.7039

Table 3: Evaluation results of different LVLMs on the MSCOCO-Cap dataset

annotators via Amazon Mechanical Turk[1] and pay 15-20 USD per hour. The average time to complete all steps of the annotation process is 212.8 seconds. More details about the annotation process are provided in the Appendix.

4.2 Correlations with Different Metrics

To verify the superiority of our proposed atomic metric, we compare it with several generation metrics: BLEU-{1-4} (Papineni et al., 2002), Rouge-{1,2, L} (Lin, 2004), METEOR (Banerjee and Lavie, 2005b), CHAIR (Rohrbach et al., 2018), and CLIP-Score (Hessel et al., 2021). To ascertain the reliability of human evaluation, we determined the Fleiss' kappa values across all annotators for the sub-sentence identification task, arriving at a value of 67.5%. This signifies a robust consensus among the annotators. Additionally, for the definitive faithfulness score, we derived Fleiss' kappa values involving all annotators and achieved a result of 47.7%. This suggests a moderate level of concordance among the evaluation participants.

Table 1 delineates the correlation between various evaluation metrics and human judgment regarding LVLM hallucinations, gauged using Spearman's ρ and Kendall's τ. One particularly intriguing observation is the CHAIR metric, which exhibits a pronounced negative correlation, even though it was specifically engineered for object hallucination evaluation. 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 scope may not adeptly address fine-grained and open-domain hallucinations, thus diminishing its efficacy and resonance with more comprehensive human evaluations. However, amid the varied metrics landscape, our metric MULTIMODAL FAITHSCOREdistinctly shines. It registers a robust positive correlation, emphasizing its superior alignment with human perceptions.

5 Evaluating Hallucinations with MULTIMODAL FAITHSCORE

5.1 Models

We selected six open-source widely used large vision-language models for evaluation. 1) MiniGPT-4 (Zhu et al., 2023); 2) LLaVA (Liu et al., 2023b); 3) InstrucBLIP (Dai et al., 2023); 4) Multimodal-GPT (Gong et al., 2023); 5) mPLUG-Owl (Ye et al., 2023); 6) LLaVA-1.5 (Liu et al., 2023a). In particular, these LVLMs are composed of three essential components: a visual encoding module, an alignment mechanism, and a large language model. Furthermore, all of these models have undergone fine-tuning using curated datasets of visual instruction data.

5.2 Datasets

To assess the performance of existing LVLMs, we conducted experiments using various datasets. Here is a description of each dataset: (1) MSCOCO-Cap: This dataset is designed for the image captioning task. We randomly selected 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:

[1] https://requestersandbox.mturk.com/.

Model	Conversation	Detailed Description	Complex Question	Overall
MiniGPT-4	0.5248	0.5042	0.5169	0.5153
LLaVA	0.4562	0.4205	0.4236	0.4334
LLaVA-1.5	0.4763	0.4708	0.4793	0.4755
InstructBLIP	0.4952	0.4813	0.4762	0.4842
Multimodal-GPT	0.4184	0.4034	0.4371	0.4197
mPLUG-Owl	0.5131	0.5202	0.5426	0.5253

Table 4: Sentence-level evaluation results of different LVLMs on the LLaVA-1k dataset.

Model	Performance
MiniGPT-4	0.4024
LLaVA	0.3298
LLaVA-1.5	0.4034
InstructBLIP	0.4845.
Multimodal-GPT	0.5364
mPLUG-Owl	0.3716

Table 5: Sentence-level evaluation results of different LVLMs on the MSCOCO-Cap dataset

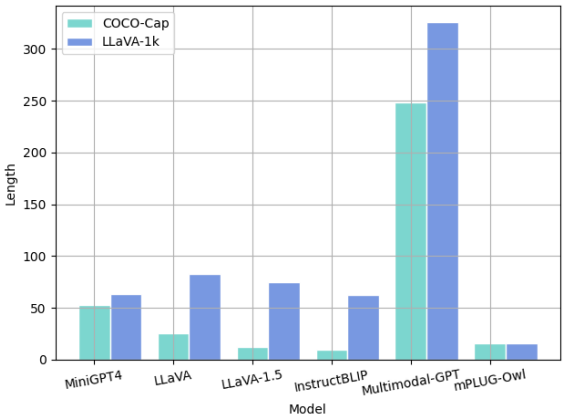

Figure 4: The illustration of the length distribution over different models and datasets.

Similar to the LLaVA dataset, we extracted 1,000 images from the COCO validation set and generated three sample types (*i.e.,* detailed description, conversation, and complex question) for each image, following the data augmentation methodology outlined in (Liu et al., 2023b).

5.3 Hallucination Evaluation

Table 2 and Table 3 present a comprehensive performance comparison of various models in terms of MULTIMODAL FAITHSCOREwhen benchmarked on the LLaVA-1k and MSCOCO-Cap datasets. (1) m-PLUG-Owl distinctly outperforms its counterparts on the LLaVA-1k and MSCOCO-Cap datasets. This demonstrates its preeminent capability in achieving and maintaining faithfulness during generation processes. A significant contributor to the model's standout performance is its extensive utilization of visual instruction data. This implies that leveraging vast amounts of visual data might be key for future improvements in the field. (2) It's worth noting that different models exhibit varied strengths across tasks. For instance, while mPLUG-Owl leads in the "Conversation" and "Detailed Description" tasks, InstructBLIP outshines in the "Complex Question" category. This may be because of the differences between data that is Used to do instruction tuning. (3) Compared to LLaVA, LLaVA-1.5 demonstrates significantly superior performance across various tasks. This improvement underscores the effectiveness of incorporating additional academic task-related data and leveraging advanced cross-modal connectors for multimodal fusion.

5.4 Sentence-level Hallucination

To further understand the faithfulness of LVLMs, we evaluate them with the proposed sentence-level MULTIMODAL FAITHSCORE. Table 4 and Table 5 show the sentence-level evaluation across different LVLMs. Upon analyzing the performance of different LVLMs across multiple datasets, several insights emerge: (1) The MiniGPT-4 model consistently ranks among the top-performing models in most categories across LLaVA-1k and MSCOCO-Cap datasets. The result indicates the consistency of the proposed metric across datasets to an extent. (2) While mPLUG-Owl has achieved successful performance in MULTIMODAL FAITHSCORE, it performs less favorably in terms of sentence-level hallucination evaluation. This is understandable because hallucinations may be dispersed loosely and scattered throughout the sub-sentences. (3)

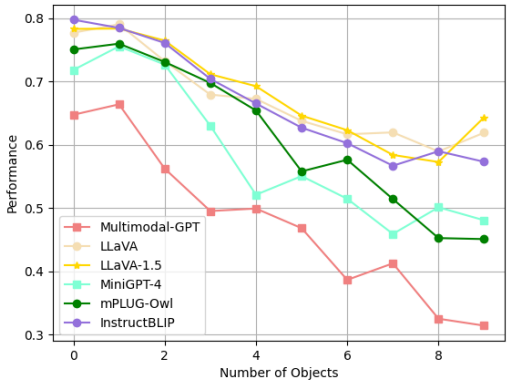

Figure 5: MULTIMODAL FAITHSCOREvs. numbers of objects (*i.e.,* entities) in the answer generated by LVLMs on LLaVA-1k. As the amount of entities increases the performance (*i.e.,* MULTIMODAL FAITHSCORE) begins to decline.

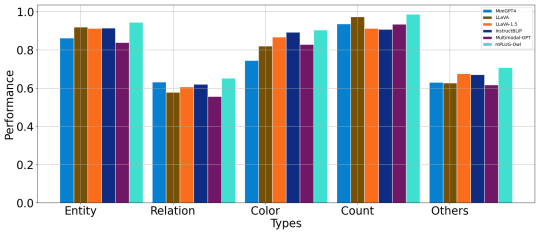

Figure 6: MULTIMODAL FAITHSCOREon each type of atomic facts in the LLaVA-1k benchmark. The types are "Entity", "Relation", "Color", "Count", and "Others".

MiniGPT-4 and Multimodal-GPT appear to be strong contenders for many multimodal tasks. This is different compared with the results in Section 5.3. The potential reason may be that the hallucinations generated by MiniGPT-4 and Multimodal-GPT are very concentrated, and they tend to appear densely in one or a few sentences. (4) Similar to the results in Section 5.3, LLaVA-1.5 further advances the performance of LLaVA across various tasks.

5.5 Other Analysis

The Influence of Answer Length on Hallucination. To further elucidate the impact of answer length on model-generated hallucinations, we analyzed answer lengths across various LVLMs on different datasets. As illustrated in Figure 4, 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, which could explain its enhanced fidelity across various tasks. Interestingly, the captioning task, despite having the shortest average answer length among all testing tasks, showed lower faithfulness in generated content. This may be attributed to the fact that captioning sentences mainly are visual descriptions, placing a heightened emphasis on the model's accuracy and faithfulness.

The influence of multiple objects. Figure 5 shows how the number of entities (*i.e.,* objects) in the answer generated by different models affects the MULTIMODAL FAITHSCOREscore. According to this figure, it is evident that the model's faithfulness varies with the number of objects. While all models start off with relatively high scores, their performance generally declines as the number of objects increases. For example, Instruct-BLIP starts with a high of 0.798 for 1 object and sustains a relatively low score of 0.573 for 10 objects.

Analysis on Types of Hallucination When comparing the performance metrics of various models across different categories, we can deduce the respective strengths and potential vulnerabilities of each in maintaining faithfulness. From Figure 6, we can observe that while mPLUG-Owl consistently excels across most categories, other models also showcase strengths in specific domains. For instance, Multimodal-GPT performs notably well in the color and count categories. The varied performance across categories underscores that most models exhibit differential strengths. However, achieving consistently high faithfulness across a diverse range of categories remains a formidable challenge for LVLMs.

6 Conclusion

In this paper, we introduce a novel metric called MULTIMODAL FAITHSCOREfor evaluating free-form answers generated by LVLMs. Compared to previous metrics, MULTIMODAL FAITHSCOREoffers a finer level of granularity, interpretability, and closer alignment with human judgments. Our quantitative analysis demonstrates that current LVLMs continue to grapple with the hallucination problem. We anticipate that MULTIMODAL FAITHSCOREwill prove invaluable for evaluating forthcoming advanced LVLMs.

Limitations

It's worth noting that, at present, MULTIMODAL FAITHSCOREelies on ChatGPT, which can be computationally expensive. Therefore, in the future, researchers can implement this metric using open-source models to make it more accessible and widely applicable.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

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

Satanjeev Banerjee and Alon Lavie. 2005b. 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 2005, Ann Arbor, Michigan, USA, June 29, 2005*, 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 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

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 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 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 Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

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 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1642–1661. Association for Computational Linguistics.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *CoRR*, 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 *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning.

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

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 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*, 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. 2023a. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation.

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

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 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, 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 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4035–4045. Association for Computational Linguistics.

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.

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.

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.

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.

A Experimental Detail

We infer all VLMs for answer generation on an NVIDIA A100 GPU. We fix the generation temperature for all VLMs to get a stable result.

B Likert Scale Guideline

Specifically, suppose the generated answer where a generated answer comprises n atomic facts, out of which x are designated as hallucinations. Both n and x are tallied 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 $x = 0$.

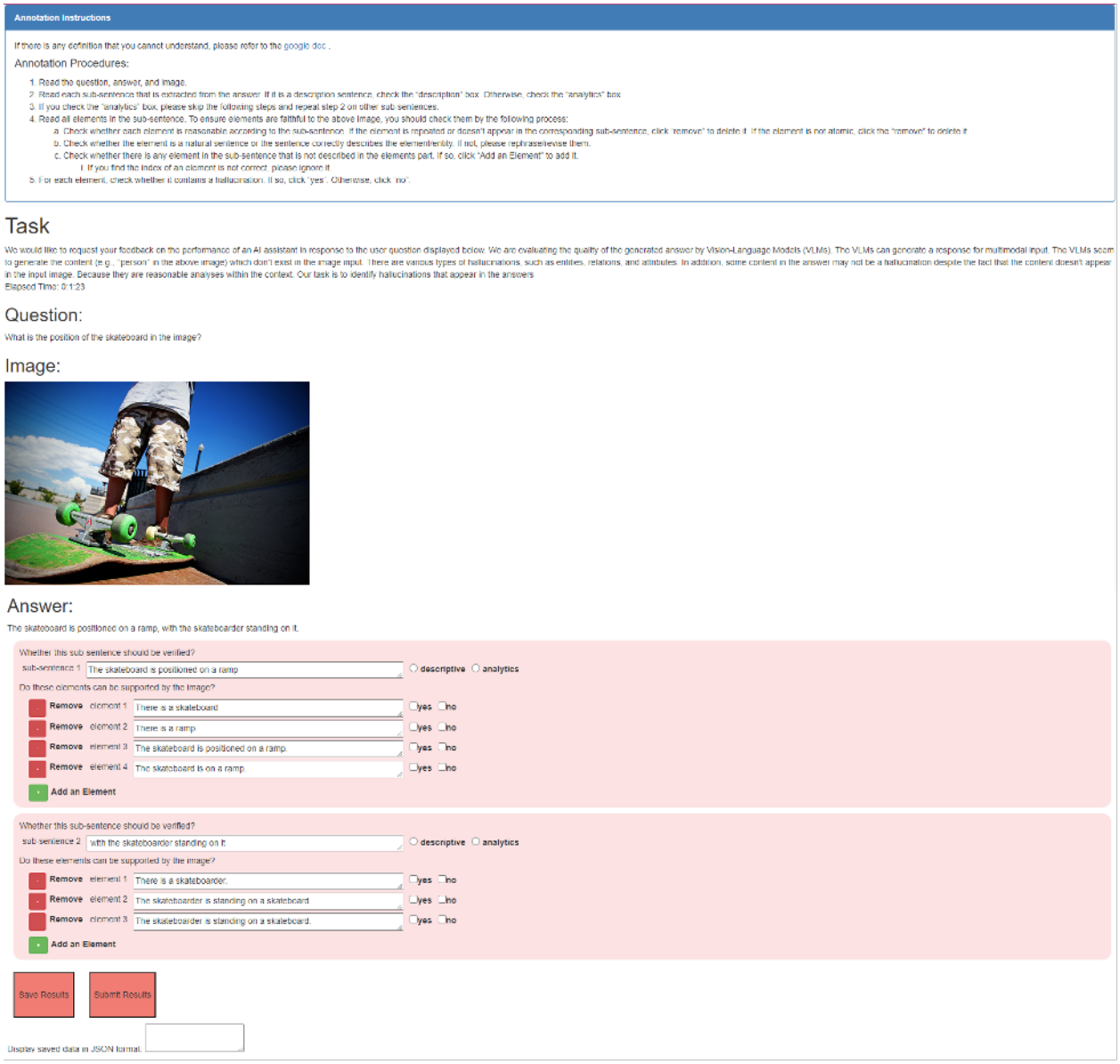

Figure 7: System software User Interface (UI) for annotators.

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 8: Instructions for data annotation.

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] denots 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 9: Prompt for sub-sentence identification.

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 10: Prompt for atomic fact generation .