Defining and Quantifying Visual Hallucinations in Vision-Language Models

Vipula Rawte[1]*, Aryan Mishra[2], Amit Sheth[1], Amitava Das[1]
[1]AI Institute, University of South Carolina, USA
[2]IISER, Bhopal
{vrawte}@mailbox.sc.edu

Abstract

The troubling rise of *hallucination* presents perhaps the most significant impediment to the advancement of responsible AI. In recent times, considerable research has focused on detecting and mitigating hallucination in Large Language Models (LLMs). However, it's worth noting that hallucination is also quite prevalent in Vision-Language models (VLMs). In this paper, we offer a fine-grained discourse on profiling VLM hallucination based on the *image captioning* task. We delineate eight fine-grained orientations of visual hallucination: i) *Contextual Guessing*, ii) *Identity Incongruity*, iii) *Geographical Erratum*, iv) *Visual Illusion*, v) *Gender Anomaly*, vi) *VLM as Classifier*, vii) *Wrong Reading*, and viii) *Numeric Discrepancy*. We curate **V**isual **H**allucI**n**ation e**L**ici**T**ation (VHILT), a publicly available dataset comprising 2,000 samples generated using eight VLMs across the image captioning task along with human annotations for the categories as mentioned earlier. To establish a method for quantification and to offer a comparative framework enabling the evaluation and ranking of VLMs according to their vulnerability to producing hallucinations, we propose the *Visual Hallucination Vulnerability Index (VHVI)*. In summary, we introduce the VHILT dataset for image-to-text hallucinations and propose the VHVI metric to quantify hallucinations in VLMs, targeting specific visual hallucination types. A sample is available at: https://huggingface.co/datasets/vr25/vhil.

Contributions

- Identification of Hallucination Categories: The paper identifies and categorizes various types of visual hallucinations in 8 VLMs. These include 8 categories listed in figure 1 and section 1.
- Creation of Visual Hallucination Dataset (VHiLT): The dataset comprises 2000 samples using 8 contemporary VLMs. Human annotations for the identified categories are included as well (section 2).
- Visual Hallucination Vulnerability Index (VHVI): We propose an evaluation metrics VHVI for quantifying and comparing the vulnerability of VLMs to produce hallucinations (section 3). This index is designed to serve as a tool for evaluating and ranking VLMs, contributing to the ongoing discourse on policy-making to regulate AI development.

1 Visual Hallucination - an extensive categorization

Despite the rapid advances in Generative AI, policymakers (Janjeva et al.) are primarily concerned with the issue of hallucinations. These occurrences of *hallucinations* pose a significant risk of eroding trust in technology. For instance, when Google's Bard AI "hallucinated" during its initial public demonstration, Alphabet experienced a temporary loss of $100 billion in market value (Olson, 2023).

The study of hallucinations for LLMs has recently attracted considerable attention (Rawte et al., 2023; Tonmoy et al., 2024). This paper delves into visual hallucination, a phenomenon notably prevalent in numerous recent VLMs. Given that this field is still emerging, it is imperative to initially comprehend, classify, and quantify these phenomena while establishing a benchmark. This will aid the scientific community in collectively addressing this issue. Compared to recent research (Huang et al., 2024; Liu et al., 2024; Fieback et al., 2024), which has primarily investigated object hallucination in VLMs using limited data. This paper aims to provide a comprehensive categorization of VLM hallucinations. We defined eight categories of Visual Hallucination:

➠ **Contextual Guessing (CG)** `alarming` When the model generates unrelated elements that bear no resemblance to the subject at hand, highlighting the non-deterministic nature of the model.

*Corresponding author.

Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025), pages 501–510
May 3, 2025

➠ **Identity Incongruity (II)** `alarming` It's when the model can't differentiate between a person's real and fake identity traits, causing a mismatch with the predicted identity.

➠ **Geographic Erratum (GE)** `alarming` In this scenario, the model produces an inaccurate prediction or guess related to the geographical location or landmark of the place under consideration.

➠ **Visual Illusion (VI)** `mild` The model can be misled, creating a distorted perception that deviates from reality, causing the model's output to be partially inaccurate due to a specific aspect of the image.

➠ **Gender Anomaly (GA)** `mild` The model provides an inaccurate representation of gender identity.

➠ **VLM as Classifier (VC)** `mild` This is a situation where the model's proficiency is assessed based on its ability to differentiate between two/more entities.

➠ **Wrong Reading (WR)** `low` When a text is engraved in an image, but the VLM read it wrong.

➠ **Numeric Discrepancy (ND)** `low` When the model encounters difficulty accurately counting the number of entities within the analyzed image leading to an inaccurate count.

Since VLMs focus on image captioning, our VHILT dataset integrates both. Unlike prior studies (Huang et al., 2024; Liu et al., 2024; Fieback et al., 2024) with limited data, we provide the most comprehensive dataset and classification of visual hallucinations.

Caption hallucination in VLMs, or object hallucination, occurs when descriptions misrepresent an image or omit key details (Fig. 1). Studies (Biten et al., 2022; Li et al., 2023; Zhou et al., 2023) link this to co-occurrence, uncertainty, misalignment between visual and language annotations (Zhai et al., 2023), inadequate training (Chen et al., 2023b), and language bias (Guan et al., 2023). While its causes remain debated, the issue's prevalence highlights the need for further research.

2 VHILT dataset

The rise of Generative AI has fueled online misinformation, as highlighted by the EU (Commission, 2022). To address this, we focus our visual hallucination dataset on the news domain. Since accurate annotations require factually correct references, we use the *New York Times Twitter handle* (NewYorkTimes, 2024) as our trusted source, covering a decade (2011–2021) of multimodal data. NYT tweets, authored by professional journalists, ensure grammatical accuracy and avoid common Twitter issues.

We specifically selected image-containing tweets for studying visual hallucinations, applying rigorous filtering to remove duplicates, irrelevant content, non-English tweets, hashtags, and URLs, retaining only original, relevant alphanumeric data.

2.1 Choice of VLMs: Rationale and Coverage

We selected SoTA VLMs for image captioning, including Kosmos-2 (Peng et al., 2023), MiniGPT-V2 (Chen et al., 2023a), and Sphinx (Lin et al., 2023). Appendix A details our selection criteria. As the field evolves, VHILT benchmark leaderboards will remain accessible for ongoing research.

2.2 Caption hallucination

We used NYT news images and fed them into Kosmos-2, MiniGPT-V2, and Sphinx to generate text captions. At this point, we have the image, caption generated by VLMs, and the actual tweet aka news headline associated with the image obtained from NYT. We also have bounding boxes and grounding information obtained from the VLMs. We provided all this information to our in-house annotators and asked them two questions: i) *Do you observe any visual hallucinations in this VLM-generated caption? Please annotate it at the sentence level*. It's worth noting that text captions may contain multiple sentences. ii) *If there is a visual hallucination, could you please describe its type?* Four in-house annotators were involved in the annotation process. After annotating 2000 instances, they collectively discussed and finalized the eight categories.

We report Fleiss's kappa (κ) (Fleiss's_Kappa) and Krippendorff's alpha (α) (Krippendorff's_Alpha) scores (see table 1) to access the reliability of agreement between the four annotators[1].

In summary, we observed two key points: i) There are instances where two or more hallucination categories are present, leading to confusion among annotators. We deliberately avoided multi-

[1] Four student interns

Figure 1: An illustration of hallucination across your multiple categories. Here, we have used VLMs like KOSMOS-2(Peng et al., 2023), MiniGPT - v2(Chen et al., 2023a), Sphinx(Lin et al., 2023) to generate captions, and the text in red color represents the particular word that is hallucinating and an added line for explanation.

	Fleiss's kappa	Krippendorff's alpha
Is hallucinated?	0.8211	0.846
Category	0.7846	0.8499

Table 1: Inter-Annotator Scores for captioning task across categories and hallucination detection.

class classification at this point; ii) Additionally, we identified new types of hallucination beyond the eight prevalent categories. We intentionally excluded such instances with skewed categorical examples, as we believe they are rare cases, and our focus is on investigating prevalent visual hallucination categories.

2.3 Annotation Process

To maintain high-quality data annotation, we conducted in-house annotation on a small portion of the data. We conducted an extensive in-house study to categorize visual hallucinations, annotating 2,000 samples for image captioning task.

3 Visual Hallucination Vulnerability Index (VHVI)

As VLM usage grows, their tendency to hallucinate lacks standardized evaluation. To address this, we introduce VHVI, a comparative spectrum for ranking VLMs by hallucination susceptibility, specifically in image captioning.

When defining VHVI, we take several factors into account. Firstly, not all captions/answers generated by a VLM are hallucinated, so it is important to determine the ratio of actual hallucinated captions/answers with the total number of captions/answers. In this context, we consider U as the total number of captions/answers produced by a VLM. Moreover, VLMs can exhibit varying degrees of hallucination, including alarming, mild, and low types. For instance, if we have two VLMs and their total number of generated hallucinations in terms of captions/answers are the same, but VLM1 produces significantly more alarming hallucinations than VLM2, we must rank VLM1 higher in terms of VHVI. This comparative measure is achieved using multiplicative damping factors, δ_H, δ_M, and δ_L which are calculated based on $\mu \pm rank_x \times \sigma$. Initially, we calculate the HVI for all the VLMs, considering δ_H, δ_M, and δ_L as one. With these initial VHVIs, we obtain the mean (μ) and standard deviation (σ), allowing us to recalculate the HVIs for all the LLMs. The resulting HVIs are then ranked and scaled, providing a comparative spectrum as presented in equation 1, similar to z-score normalization (Wikipedia_zscore) and/or min-max normalization (Wikipedia_min_max). Having damping factors enables easy exponential smoothing with a handful of data points, 3/5 in this case. Finally, for ease of interpretability, VHVI is scaled between $0 - 100$. Please see figure 2 for the VHVI ranking of three VLMs.

3.1 VHVI captioning

When calculating $VHVI_{capt}$, we take into account the probability of each visual hallucination category. For example, H_{CG}^{C} represents the total number of instances of Contextual Guessing out of the total U generated captions. Therefore, the probability of this VLM generating Contextual Guessing-type hallucination is (H_{CG}^{C}/U).

$$VHVI_{capt} = \frac{100}{U}[\textstyle\sum_{x=1}^{U}(\delta_H * (H_{CG}^{C} + H_{II}^{C} + H_{GE}^{C}))+ (\delta_M * (H_{VI}^{C} + H_{GA}^{C} + H_{VC}^{C})) + (\delta_L * (H_{WR}^{C} + H_{ND}^{C}))] \quad (1)$$

VLM	Size	VHVI (0-100)
Kosmos-2	1.6B	54 -
MiniGPT-v2	7B	48 -
Sphinx-1k	13B	39 -

Figure 2: VHVI for VLM models based on captioning task using equation 1. The model size is found to be inversely proportional to VHVI.

Implications derived from VHVI

- Alarming hallucination categories, such as contextual guessing, identity incongruity, geographic erratum, and visual illusion, are prevalent in VLMs beyond a specific size. For instance, Kosmos-2 for image captioning is more vulnerable to these categories of hallucination.
- The numeric discrepancy, wrong reading, and VLM as a classifier are pervasive issues across all VLMs across both tasks.

4 Conclusion

The enthusiasm and achievements surrounding Generative AI models have led to their widespread adoption, and this trend is only expected to flourish. However, one of the most significant challenges faced by these models today is hallucination. In light of this, the benchmark and *Visual Hallucination Vulnerability Index (VHVI)* will continue to serve the wider scientific community and aid policy-makers. VHILT benchmark and VHVI will be publicly open for further collaborative updates.

5 Limitations

On June 14th, 2023, the European Parliament successfully passed its version of the EU AI Act (European-Parliament, 2023). Following this, many other countries began discussing their stance on the evolving realm of Generative AI. A primary agenda of policymaking is to protect citizens from political, digital, and physical security risks posed by Generative AI. While safeguarding against misuse is crucial, one of the biggest concerns among policymakers is the occurrence of unwanted errors by systems, such as hallucination (Janjeva et al.). We firmly believe that the proposed VHVI can provide valuable insights for policymakers, enabling them to make informed decisions. As we make VHVI publicly available, we are confident that it will garner attention within the scientific community. We anticipate that researchers will utilize VHVI to evaluate various VLMs, contributing to further advancements in this field.

Limitations: In this paper, we introduce an exclusive and comprehensive benchmark dataset for hallucination, named VHILT. We propose hallucination across the main task: Image Captioning, each further divided into eight categories. Additionally, we map these categories with the degree, i.e., alarming, mild, and low. We think paying close attention to the following aspects in future efforts is essential.

Limitation 1: To keep things simple, we annotated only one category per sentence in the captioning task, even though we recognized the existence of instances with multiple classes and labels. For instance, in the example (see figure 3), there are two kinds of hallucination, namely Numeric Discrepancy and Gender Anomaly, present in the shown Example. Although how minuscule the problem seems to be, but the probability of encountering such blends of hallucinations isn't completely zero. Therefore it is important to resolve this issue for the betterment of VLMs, so we want to explore this direction in the immediate future.

Limitation 2: In this research, we have selected 8 VLMs. Given the ever-evolving nature of VLM development, new models are continually emerging, and we recognize that our choice may not cover all the available options. Considering this, we intend to make the VHILT benchmark and the *VHVI* openly available for collaborative updates and contributions.

Limitation 3: Another limitation worth noting is VLMs continuously evolve, so the results may change if tried at a later time, as described in figure 4; nevertheless, our results in open source will continue to provide insight.

Question: Identify the Gender of the people, in the order they are positioned.
Answer Generated by MiniGPT-4: The people in the image are all female.

Question: How many people are there ?
Answer Generated by MiniGPT-4: There are four people in the image.

Explanation: Firstly, there are five people in the image, and secondly out of five two are men and rest being women.
Category : Wrong Reading and Numeric Discrepancy

Figure 3: Example exhibiting both Gender Anomaly and Numeric Discrepancy category of hallucination. Since there were Five people, but the model(MiniGPT-4) Identified only Four, also every one of them has been identified as female, even though there were male counterparts.

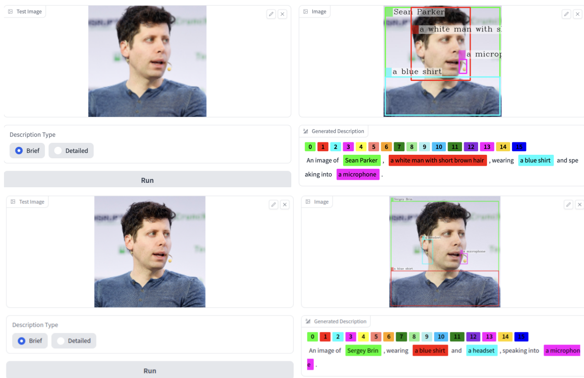

Figure 4: Example of *Sam Altman* being predicted as *Sean Parker* when the model (Kosmos-2) is run for the first time and *Sergey Brin* for the second time.

6 Ethics Statement

Through our experiments, we have uncovered the susceptibility of VLMs to hallucination. In developing VHVI, we intend to provide a framework that can inform future research and policies in this domain. However, we must address the potential misuse of our findings by malicious entities who may exploit AI-generated images, such as creating indistinguishable fake news from human-written content. We vehemently discourage such misuse and strongly advise against it.

References

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023b. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*.

European Commission. 2022. Eu code of conduct against online hate speech: latest evaluation shows slowdown in progress.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*.

European-Parliament. 2023. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

Laura Fieback, Jakob Spiegelberg, and Hanno Gottschalk. 2024. Metatoken: Detecting hallucination in image descriptions by meta classification. *arXiv preprint arXiv:2405.19186*.

Fleiss's_Kappa. Fleiss's kappa.

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. *arXiv preprint arXiv:2305.04790*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu Ruiqi Xian Zongxia Li, Xiaoyu Liu Xijun Wang, Lichang Chen Furong Huang Yaser Yacoob, and Dinesh Manocha Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv e-prints*, pages arXiv–2310.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multimodal large language models. *arXiv preprint arXiv:2402.14683*.

Ardi Janjeva, Alexander Harris, Sarah Sarah, Alexander Kasprzyk, and Anna Gausen. The rapid rise of generative ai.

Krippendorff's_Alpha. Krippendorff's alpha.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024. Phd: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*.

NewYorkTimes. 2024. The new york times twitter. `https://twitter.com/nytimes`.

Emily Olson. 2023. Google shares drop $100 billion after its new ai chatbot makes a mistake.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Tonmoy, Aman Chadha, Amit P Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models–an extensive definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2310.04988*.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models.

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

Wikipedia_min_max. Normalization.

Wikipedia_zscore. Normalization.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. 2023. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pages arXiv–2310.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Appendix

This section provides additional examples to assist in the understanding and interpretation of the research work presented.

A Details on Choice of VLMs: Rationale and Coverage

We shortlisted five SOTA models for VQA InstructBlip(Dai et al., 2023), MiniGPT - v2(Chen et al., 2023a), Multimodal-gpt(Gong et al., 2023), LLava(Liu et al., 2023), mPlug-Owl(Ye et al., 2023). Recent work on visual hallucination in VLMs chooses these models LURE(Zhou et al., 2023), POPE(Li et al., 2023), and HaELM(Wang et al., 2023) for analysis. In a similar line of reasoning for the captioning task, we shortlisted three SOTA models for studying hallucination in captioning, namely Kosmos-2(Peng et al., 2023), MiniGPT-v2(Chen et al., 2023a), and SPHINX(Lin et al., 2023).

B Additional Examples for Captioning

In the following, we provide additional examples of captioning hallucination generated by three models.

B.1 Additional Examples for captioning using Kosmos-2

B.2 Additional Examples for captioning using MiniGPT-V2

B.3 Additional Examples for captioning using Sphinx

Figure 5: Examples from captioning task using KOSMOS-2

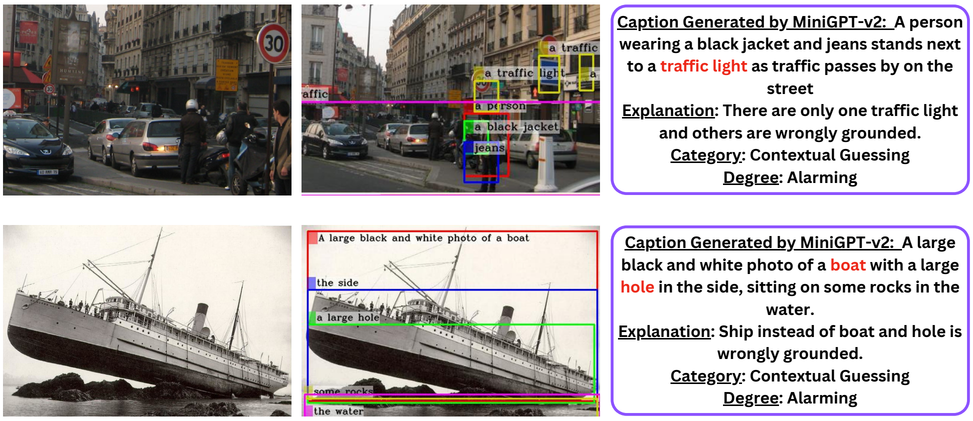

Figure 6: Examples from captioning task using MiniGPT-v2

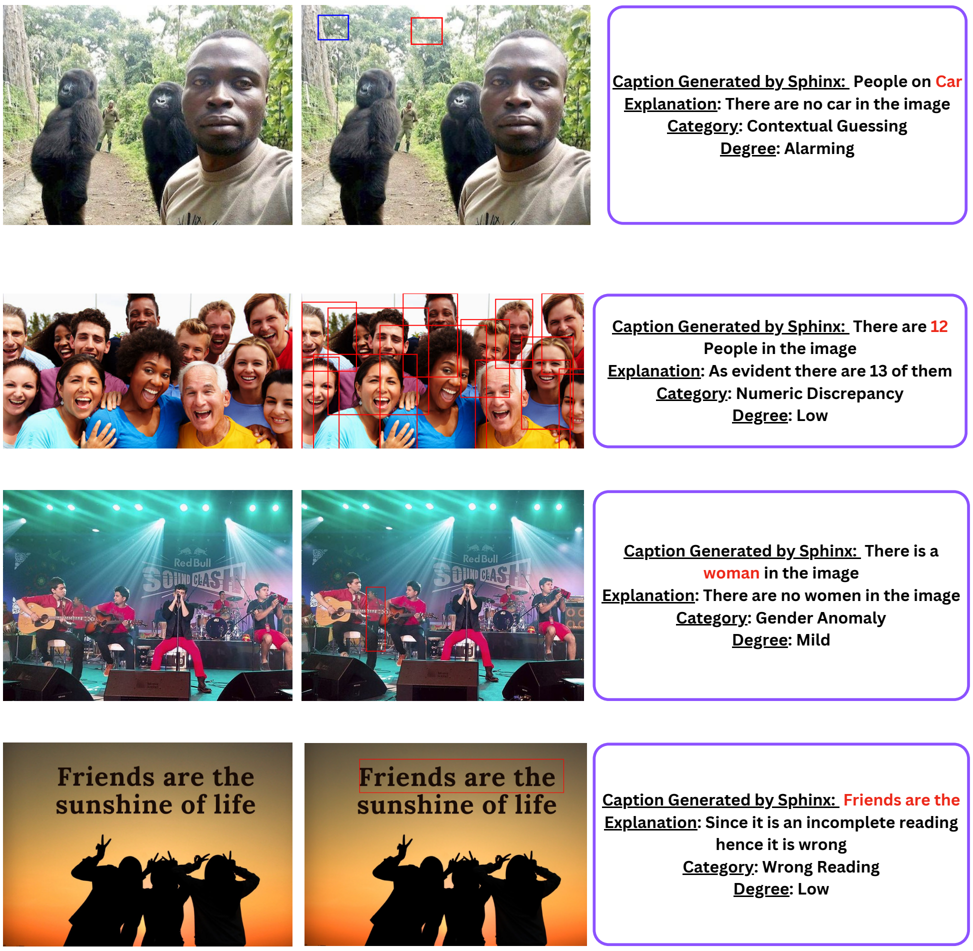

Figure 7: Examples from captioning task using Sphinx