Evaluating the Limitations of Generative Models in Image Captioning

Shang Jin Yan Sun Zehao Li Zhiqian Li
University of California, Los Angeles
{upking, yas063, zehao3, cassiezhiqq}@g.ucla.edu
Authors are listed in alphabetical order.

Abstract

Recent vision-language models (VLMs) have demonstrated remarkable performance across various tasks and domains. While most large vision-language models (LVLMs) excel at generating image captions, however, their ability to reliably distinguish correct captions from those with subtle errors remains underexplored. Unlike existing metrics for image captioning that depend on comparing reference and candidate captions, this paper introduces a more direct approach by employing Visual Question Answering (VQA) across multiple domains.

To evaluate the capabilities of these models, we address several key challenges in dataset construction: (a) sourcing questions from multiple domains, including ArtCap, FoodCap, and COCO; (b) incorporating varying levels of difficulty in multiple-choice questions; and (c) presenting images in four distinct formats: original, cropping, saturation, and with noise. We curated a 400-entries dataset **VQA4Mix** and conducted comprehensive experiments on two LVMs, LLaVa and Phi. The code for this study is publicly available at Github Repository.

1 Introduction

Image Captioning (Wang et al., 2020) refers to describing the content of an image in natural language. The task lies at the intersection of Computer Vision and Natural Language Processing. With the rapid development of encoder-decoder architectures, many pre-trained vision-language models perform well on captioning tasks (Lai et al., 2024; Rotstein et al., 2024; Wang et al., 2024), even in few-shot (Özdemir and Akagündüz, 2024) and zero-shot (Yu et al., 2024) settings. This paper brings an evaluation of LLaVa (Liu et al., 2024) and Phi (Abdin et al., 2024) models on the Visual Question Answering (VQA) task. The most widely applied image captioning metrics, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015), are based on n-gram similarity, which requires reference captions and candidate captions. As shown in Figure 1, the expression *'standing on top of a'* contributes significantly to n-gram similarity counts, which results in false positives during evaluation. These metrics have limited semantic understanding and heavily depend on the diversity and quality of the reference captions. These drawbacks inspire us to create and evaluate under a reference-free environment where the models are prompted to choose the most suitable caption from a given set of candidates.

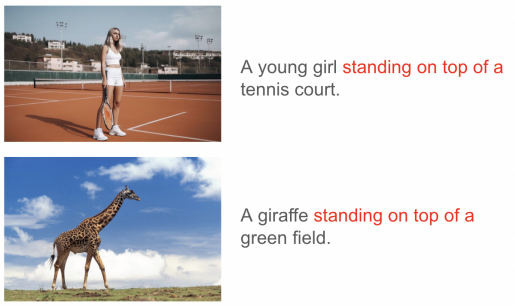

Figure 1: False Positive examples in image caption with n-gram based metrics

Motivated by these research progresses, we propose a mixed-domain Visual Question Answering dataset, **VQA4Mix**, for evaluating image captioning. This dataset covers topics in Art (Lu et al., 2024), Food (Ma et al., 2023), Cats, and People (Lin et al., 2014), varying in three difficulty levels(easy, medium, and hard) and four image patterns(original, cropping, saturation, and noise). For all datasets, the setup is as follows: the model is given an image and four candidate captions with a prompt asking it to return one answer. We evaluate two models under a zero-shot setting and measure the accuracy. Furthermore, this paper provides domain-specific performance analysis and proposes insights that could expand the range in future studies.

2 Approach and Methodology

To evaluate Vision-Language Models (VLMs), we selected 100 images from various domains, including the *cat* and *people* subcategories from the COCO dataset (Lin et al., 2014), as well as samples from FoodCap (Ma et al., 2023) and ArtCap (Lu et al., 2024). Leveraging these well-known image captioning datasets, we developed a comprehensive evaluation framework with the following components:

- **MCQ Questions:** Multiple-choice questions designed for VLMs to select the correct answer.
- **Image Augmentation:** Modifications such as adjusting image size and color to evaluate model robustness.
- **Unique Prediction Accuracy Metrics:** Custom evaluation metrics enabling comparative analysis of model performance across different tasks.

2.1 MCQ Task Design

We designed three levels of multiple-choice (MCQ) questions: **easy**, **medium**, and **hard**. GPT-4 was used as a baseline (OpenAI, 2023) to generate questions, each with four answer choices related to image captions for specific images. Among the four options, only one correct answer is derived from the ground truth annotations in the corresponding datasets.

Figure 2: Sample image used for generating MCQ questions.

For **easy** questions, incorrect answers are intentionally obvious yet loosely related to the image content.

In **medium** questions, distractors are contextually relevant but contain inaccurate details or syntactically flawed language.

A. A man cleaning a large stainless steel refrigerator.
B. A woman placing a tray into a small microwave.
C. A man opening up a big stainless steel oven.
D. A child closing the door of a wooden cabinet.

Figure 3: Example of an easy MCQ question.

A. A woman closing a small white microwave.
B. A man standing next to a wooden cupboard.
C. A man opening up a big stainless steel oven.
D. Someone cooking on a gas stove.

Figure 4: Example of a medium MCQ question.

Finally, in **hard** questions, incorrect answers are closely related to the context and designed to be challenging, potentially confusing both human evaluators and VLMs without careful observation and thorough understanding of the image.

A) A man cleaning a big stainless steel oven.
B) A man closing a big stainless steel oven.
C) A man opening up a big stainless steel oven.
D) A man cooking with a big stainless steel oven.

Figure 5: Example of a hard MCQ question.

By providing these three difficulty levels, we aim to comprehensively evaluate the performance of VLMs, offering interpretable and granular assessment results.

2.2 Image Augmentation

In order to further analyze the capabilities of Vision-Language Models (VLMs), we applied a series of image augmentation techniques, including cropping, saturation adjustment, and the addition of noise to the original images. These augmentations were designed to evaluate the models' ability to interpret images with varying levels of detail, altered

color dynamics, and degraded visual quality.

- **Original Image:** The unmodified image serves as the baseline, allowing us to assess the model's performance under normal conditions.
- **Cropped Image:** By removing portions of the image, we test the model's ability to reason with incomplete visual information and infer details from partial context.
- **Saturation Adjustment:** Altering the image's color intensity helps evaluate the model's robustness to variations in color representation and its capability to adapt to different visual environments.
- **Noisy Image:** Adding visual noise simulates low-resolution or degraded image quality, challenging the model to extract relevant features under suboptimal conditions.

The four augmented versions of each image are depicted in Figures 6, 7, 8, and 9. By leveraging these augmentations, we aim to comprehensively evaluate the models' ability to process partial, detailed, and low-quality images, thereby providing deeper insights into their robustness and limitations.

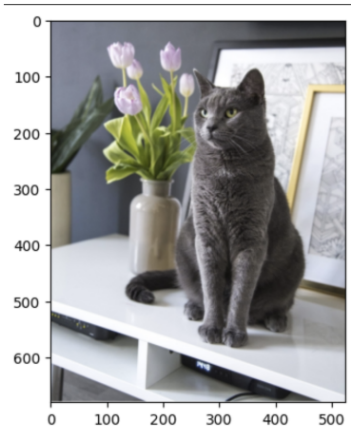

Figure 6: Original image.

2.3 Prediction Accuracy

The prediction accuracy of Vision-Language Models (VLMs) is summarized across two dimensions: **image augmentation types** and **question difficulty levels**. The rows represent four augmentation types applied to the images—*original*, *cropping*, *saturation adjustment*, and *noise addition*. The

Figure 7: Cropped image.

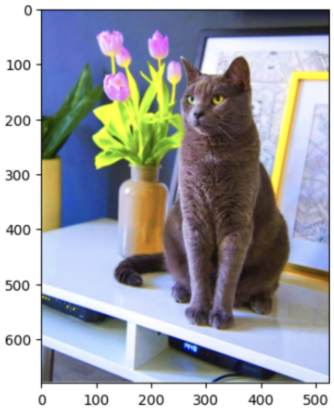

Figure 8: Image with saturation adjustment.

columns represent three levels of question difficulty: *easy*, *medium*, and *hard*.

This tabular structure enables a clear breakdown of model performance across different augmentation methods and difficulty levels, providing insights into how these factors jointly affect the models' accuracy.

3 Experiment and Result

In this section, we analyze the performance of the two Large Vision-language Models, LLava and Phi, for Visual Question Answering (VQA) across different domains, image augmentation techniques, and difficulty levels. Both qualitative and quantitative insights are presented. The experiment result is shown in Table 1.

3.1 Quantitative Analysis

3.1.1 Overall Performance

The overall performance, measured in terms of prediction accuracy, reveals that Phi outperforms LLava across all difficulty levels. For easy ques-

Domain	Augmentation	Llava - Easy	Llava - Medium	Llava - Hard	Phi - Easy	Phi - Medium	Phi - Hard
Food	Original Image	94.00	71.00	63.00	99.00	90.00	86.00
	Cropping	91.00	70.00	58.00	97.00	89.00	84.00
	Saturation	90.00	64.00	59.00	93.00	81.00	76.00
	Noise	83.00	63.00	55.00	86.00	80.00	66.00
Art	Original Image	92.00	82.00	66.00	91.00	81.00	72.00
	Cropping	88.00	76.00	64.00	87.00	76.00	70.00
	Saturation	90.00	81.00	67.00	85.00	74.00	69.00
	Noise	85.00	76.00	59.00	80.00	70.00	62.00
Cat	Original Image	98.00	93.00	88.00	100.00	97.00	90.00
	Cropping	98.00	90.00	90.00	100.00	94.00	91.00
	Saturation	98.00	89.00	59.00	99.00	96.00	88.00
	Noise	81.00	56.00	58.00	88.00	78.00	66.00
People	Original Image	100.00	89.00	81.00	98.00	93.00	90.00
	Cropping	97.00	86.00	83.00	96.00	91.00	83.00
	Saturation	100.00	88.00	83.00	98.00	91.00	83.00
	Noise	89.00	78.00	72.00	91.00	84.00	71.00

Table 1: Prediction Accuracy of Llava-v1.6-7b and Phi-3.5-vision-instruct Across Image Domains and Augmentations (Accuracy values are represented in percentages).

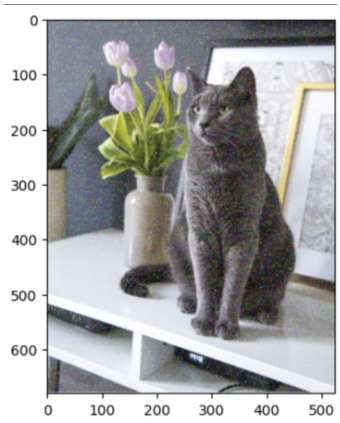

Figure 9: Image with added noise.

tions, Phi achieves an accuracy of 93.00%, slightly higher than LLava's 92.00%. For medium and hard questions, Phi's accuracy is 85.00% and 78.00%, respectively, compared to LLava's 78.00% and 69.00%. The average accuracy of Phi (85.00%) surpasses LLava (80.00%).

3.1.2 Domain-Specific Analysis

Food Images: In the food domain, Phi consistently outperforms LLava across all augmentations and difficulty levels. With original images, Phi achieves 99.00% accuracy for easy questions and 86.00% for hard questions, whereas LLava achieves 94.00% and 63.00%, respectively. Noise augmentation significantly impacts LLava's performance, with hard-question accuracy dropping to 55.00%, while Phi maintains relatively robust performance (66.00%).

Art Painting Images: For art paintings, Phi again demonstrates higher robustness. With original images, Phi achieves 91.00% for easy and 72.00% for hard questions, compared to LLava's 92.00% and 66.00%. Noise augmentation has a notable impact on both models, with LLava's hard-question accuracy reducing to 59.00%, and Phi achieving 62.00%.

Cat Images: In the cat domain, Phi performs exceptionally well, achieving 100.00% for easy and 90.00% for hard questions with original images, compared to LLava's 98.00% and 88.00%. Saturation augmentation impacts LLava's hard-question accuracy (59.00%), while Phi maintains a higher accuracy of 88.00%. Noise augmentation severely affects LLava (58.00% for hard questions), whereas Phi remains relatively stable (66.00%).

3.2 Qualitative Analysis

Impact of Augmentations: Both models exhibit reduced performance under image augmentations, with noise being the most detrimental. LLava is particularly sensitive to noise, as evidenced by significant drops in accuracy across all domains. Phi demonstrates greater resilience, maintaining higher accuracy under all augmentation techniques.

Difficulty-Level Trends: Accuracy consistently decreases as question difficulty increases, which is expected due to the increased complexity of VQA tasks. However, Phi exhibits a smaller performance gap between difficulty levels compared to LLava, suggesting its superior reasoning capabilities.

Domain-Specific Observations: Phi exhibits more consistent performance across different domains, while LLava shows more variability. For instance, LLava performs better in the cat domain than in the food and art painting domains, but its accuracy drops more drastically under augmentations.

3.3 Summary of Observations

- **Phi consistently outperforms LLava** across all domains, difficulty levels, and augmentation techniques.
- **Noise augmentation significantly impacts both models**, with LLava showing higher sensitivity.
- **Difficulty increases amplify performance gaps**, but Phi maintains better accuracy than LLava for medium and hard questions.
- **Domain-specific trends reveal Phi's robustness**, while LLava exhibits domain-dependent variability.

4 Failure Case Analysis

In this section, we analyze a failure case of a Visual Question Answering (VQA) task performed by LLava. Figure 10 shows the input image along with the ground truth caption (Solution) and the incorrect prediction made by the model.

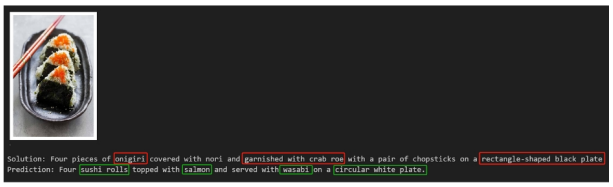

Figure 10: Failure Case Example.

4.1 Analysis of the Ground Truth and Model Prediction

The solution provided the following description: *"Four pieces of onigiri covered with nori and garnished with crab roe with a pair of chopsticks on a rectangle-shaped black plate."*

In contrast, the model predicted: *"Four sushi rolls topped with salmon and served with wasabi on a circular white plate."*

The key differences between the ground truth and the model prediction are summarized below:

- **Type of food:** The ground truth identifies the food as *onigiri*, while the model predicts *sushi rolls*.
- **Toppings:** The ground truth specifies *crab roe* as the topping, but the model incorrectly identifies it as *salmon*.
- **Accompaniments:** The ground truth makes no mention of *wasabi*, whereas the model incorrectly includes it as an accompaniment.
- **Plate shape and color:** The ground truth specifies a *rectangle-shaped black plate*, while the model predicts a *circular white plate*.

4.2 Potential Causes of the Error

The incorrect prediction by the model could be attributed to several factors:

1. **Visual Feature Misinterpretation:** The model might have misinterpreted the visual features of the food due to its reliance on pre-trained embeddings that are biased toward more common classes, such as sushi rolls and salmon, which are more frequently encountered in datasets compared to onigiri or crab roe.
2. **Dataset Bias:** Training data bias could lead to incorrect associations between visual features and semantic labels. For example, the model might have been trained on a dataset where sushi rolls and wasabi co-occur frequently, causing it to default to these predictions when the context is ambiguous.
3. **Object and Context Confusion:** The model may have struggled to distinguish the details of the plate (shape and color) and toppings due to subtle visual differences or limited resolution in the input image.
4. **Lack of Fine-Grained Understanding:** The model may lack the fine-grained understanding required to distinguish between visually similar foods, such as onigiri and sushi rolls.

4.3 Implications for Model Improvement

To address these issues, the following steps could be taken:

- Enhance the training dataset to include more diverse examples of visually similar foods, such as onigiri and sushi rolls, along with their context-specific details.
- Employ fine-tuning techniques to adapt the model to domain-specific tasks that require distinguishing subtle visual features.
- Integrate auxiliary tasks, such as object detection or scene segmentation, to improve the model's ability to analyze the context and identify specific visual elements.
- Incorporate multimodal reasoning techniques to combine textual and visual information more effectively, enabling the model to infer fine-grained details from the image.

5 Future Work

Our evaluation of Llava and Phi in Image Captioning Question Answering has provided valuable insights into their performance across different levels of task difficulty. However, there are several areas that warrant further exploration based on the analysis of failure cases:

1. **Object Recognition Accuracy:** Future investigations could focus on whether the models can accurately identify the primary objects within an image. Misidentifications in this area may lead to incorrect caption selections, particularly in cases where the main object serves as the key to determining the most appropriate caption.
2. **Attribute Understanding:** The models' ability to discern specific attributes such as shape, color, size, and texture requires deeper analysis. As shown in Section 4.1, models inaccurately distinguish the shape of sushi plate as circular. These attributes often play a critical role in distinguishing between closely related caption options, especially for medium and hard difficulty levels.
3. **Hallucinations in Outputs:** A notable failure mode involves the generation of hallucinated details that are not present in the image, such as associating sushi with wasabi when wasabi is not depicted. Future work could investigate the underlying causes of such hallucinations and propose methods to mitigate them, thereby improving the reliability of the predictions of models.

By addressing these areas, we aim to refine the evaluation framework and contribute to a deeper understanding of the limitations and potential of large language models.

6 Conclusion

Our project presented a comprehensive evaluation of two large vision-language models, Llava and Phi, in the domain of Image Captioning Visual Question Answering. By curating a dataset with varying question difficulty levels and applying diverse image augmentation techniques, we systematically assessed the models' capabilities and limitations. Our findings highlight notable differences between the two models, with Phi consistently outperforming Llava in prediction accuracy across most scenarios, including under challenging conditions such as noisy or altered images.

Key insights include the significant impact of question difficulty and image augmentation on model performance. Phi demonstrated superior resilience, maintaining higher accuracy even in the face of complex visual scenarios and subtle distinctions in question phrasing. In contrast, Llava exhibited greater sensitivity to noise and struggled more with domain-specific variations.

Failure case analyses revealed areas for improvement, including difficulties with recognizing fine-grained details, understanding object attributes, and mitigating hallucinated outputs. These challenges underscore the need for better training datasets, fine-tuning strategies, and advanced reasoning techniques to enhance model reliability and interpretability.

This work contributes to the growing body of research aimed at understanding and advancing the multimodal capabilities of generative models. By providing an evaluation framework and identifying specific limitations, We hope this work provides a useful reference for researchers seeking to better understand and improve the performance of vision-language systems in multimodal tasks.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat

Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Satanjeev Banerjee and Alon Lavie. 2005. 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.

Zhengfeng Lai, Vasileios Saveris, Chen Chen, Hong-You Chen, Haotian Zhang, Bowen Zhang, Juan Lao Tebar, Wenze Hu, Zhe Gan, Peter Grasch, et al. 2024. Revisit large-scale image-caption data in pre-training multimodal foundation models. *arXiv preprint arXiv:2410.02740*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *European conference on computer vision*, pages 740–755.

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

Y. Lu, C. Guo, X. Dai, and F.-Y. Wang. 2024. Artcap: A dataset for image captioning of fine art paintings. *IEEE Transactions on Computational Social Systems*, 11(1):576–587.

Zheng Ma et al. 2023. Food-500 cap: A fine-grained food caption benchmark for evaluating vision-language models. In *Proceedings of the 31st ACM International Conference on Multimedia*.

OpenAI. 2023. Gpt-4 technical report. https://openai.com/research/gpt-4.

Övgü Özdemir and Erdem Akagündüz. 2024. Enhancing visual question answering through question-driven image captions as prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1562–1571.

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*, pages 311–318.

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5689–5700.

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.

Haoran Wang, Yue Zhang, and Xiaosheng Yu. 2020. An overview of image caption generation methods. *Computational intelligence and neuroscience*, 2020(1):3062706.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.

Eric Yang Yu, Christopher Liao, Sathvik Ravi, Theodoros Tsiligkaridis, and Brian Kulis. 2024. Image-caption encoding for improving zero-shot generalization. *arXiv preprint arXiv:2402.02662*.