Can CLIP Count Stars? An Empirical Study on Quantity Bias in CLIP

Zeliang Zhang , Zhuo Liu , Mingqian Feng, Chenliang Xu
Department of Computer Science, University of Rochester

{zeliang.zhang, zhuo.liu, mingqian.feng,chenliang.xu}@rochester.edu

Abstract

CLIP has demonstrated great versatility in adapting to various downstream tasks, such as image editing and generation, visual question answering, and video understanding. However, CLIP-based applications often suffer from misunderstandings regarding user intent, leading to discrepancies between the required number of objects and the actual outputs in image generation tasks. In this work, we empirically investigate the quantity bias in CLIP. By carefully designing different experimental settings and datasets, we comprehensively evaluate CLIP's understanding of quantity from text, image, and cross-modal perspectives. Our experimental results reveal a quantity bias in CLIP embeddings, impacting the reliability of downstream tasks.

Figure 1: Existing models often show the quantity bias in different tasks. In this example, CLIP-based stable-diffusion model mostly generates a picture with seven pandas, while we only prompt the five.

1 Introduction

The Contrastive Language-Image Pre-Training (CLIP) model (Radford et al., 2021), trained on large-scale image-text pairs, has shown significant success in various downstream vision-language tasks, including editing (Guerrero-Viu et al., 2024; Michel et al., 2024), generation (Ganz and Elad, 2024; Liu et al., 2024), and quality evaluation (Hong et al., 2024; Deng et al., 2024). It is crucial to maintain a reliable CLIP model at the core to ensure the development of trustworthy applications built upon it (Zhang et al., 2024b).

However, several factors potentially hinder the interpretability and trustworthiness of CLIP, including the black-box nature of the learning process, uneven distributions of the training data, and the difficulty in accurately learning specific data distributions. Such issues may lead to unintended systematic errors like spurious correlations (Sagawa et al., 2020) and subgroup biases (Zhang et al., 2024a). These drawbacks not only degrade CLIP's performance in learning reliable latent representations for image and text translation, but also pose a risk of propagating unexpected biases to models that utilize CLIP for downstream tasks, thereby resulting in more challenging bugs to fix (Tanjim et al., 2024).

For instance in Fig. 1, when using Stable-Diffusion (SD) (Rombach et al., 2022), which leverages the relationship between image and text learned from CLIP (Ding et al., 2024), to generate an image of five pandas, most of the output consistently depicts seven pandas. We query the reason from the foundational CLIP model rather than the SD model at the top layer, as debugging becomes significantly more difficult and challenging if the foundational models are already biased. This intriguing phenomenon raises an important question: *Can CLIP count stars*?

To address this question, our work investigates the quantity bias in CLIP. Specifically, we empirically evaluate CLIP at two levels: uni-modal (text & image), and cross-modal interactions. For each level, we set up tasks of varying difficulty to ensure a comprehensive evaluation. Our findings highlight the need for addressing these biases to enhance the reliability and effectiveness of CLIP in real-world applications.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1081–1086

We summarize our findings as follows,

- CLIP models can not understand the concept of quantity in text-only, image-only, or cross-modality contexts.
- CLIP can distinguish the semantic difference between different quantity words but fails to compare them effectively.
- CLIP can not effectively find the semantic difference between images with different number of same or similar objects, also leading to the failure of quantity identification.

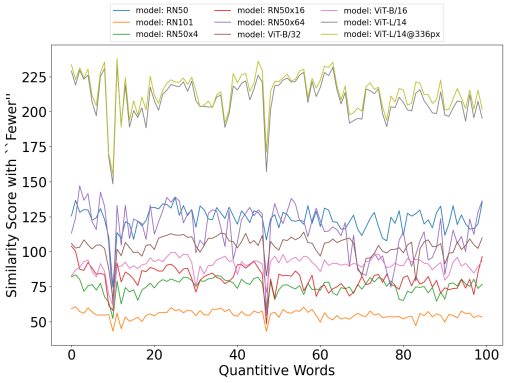

Figure 2: How "fewer" for selected different words.

2 Related work

There have been many efforts working on the model bias. Kotek et al. (2023) investigate the behavior of large language models on gender bias. Liu et al. (2022) measure the political bias in language models and propose a reinforcement learning-based method to mitigate the bias. Zhang et al. (2024a) identify the existence of subgroup bias in image classifiers and use a supervised decomposition method to discover unknown bias from the joint information from the model and inputs. Hosseini et al. (2018) find the shape bias learning by convolutional neural networks. Khayatkhoei and Elgammal (2022) discover generative models can easily learn the spatial bias from the data. Heinert et al. (2024) and Hönig et al. (2024) research on the texture bias of deep learning models and downstream tasks.

Different from previous task-specific and application-driven studies on model bias, we study the bias of the embedding in CLIP. There are mainly two reasons motivating us: First, various studies have identified that there is fruitful semantic information in the embedding, where the existence of bias could have a great impact on the whole model. Second, as CLIP serves as a vision-language foundation model in many downstream tasks and model developments, it can help us understand the model behavior by studying the bias issue of the used foundation model. In this work, we study the existence of quantity bias from the CLIP embedding level for a better understanding of the failure of generative models.

3 Quantity Bias in CLIP

3.1 Experiment design

Overview. We study the quantity bias in CLIP at two levels: uni-modal and multi-modal. In each modal capacity, we examine the concept association between different quantitative nouns and comparative descriptions. Then, we investigate the quantity bias in cross-modal capacity. Furthermore, we use CLIP-based retrieval results to reflect the impact of quantity bias in CLIP.

Models. Nine CLIP models are evaluated in our study, including RN-50 (He et al., 2016), RN-101, RN-50x4, RN50x16, RN50x64, ViT-B/32 (Han et al., 2022), ViT-B/16, ViT-L/14, and ViT-L/14@336px. We get the pre-trained models from the CLIP library (Radford et al., 2021).

Datasets. We manually construct the dataset for the quantity bias study. For the text modality, we create various quantity-related nouns such as 'zero,' 'three,' and 'hundreds.' For the image modality, we generate images with different numbers of circles, where the positions of the circles are randomly sampled. To better reflect CLIP's knowledge of quantity, we use descriptive nouns to benchmark various detailed quantity nouns, such as 'many,' 'fewer,' and 'lots of.' We study the quantity bias at the embedding level, which encodes the rich semantic information learned by the CLIP model.

Evaluation metric. Following previous studies, we use the inner product as the similarity score between embedding to evaluate the semantic correlation between different words and concepts.

3.2 Evaluation on the uni-modal capacity

Texts are mostly used in CLIP-based downstream tasks, which are more intuitive for human understanding. While our human can easily describe what is less and what is more, here, we question whether the CLIP also knows.

We build a small dataset containing 25 specific quantity nouns, ranging from 0 to 100. Two benchmark words for comparable description are used:

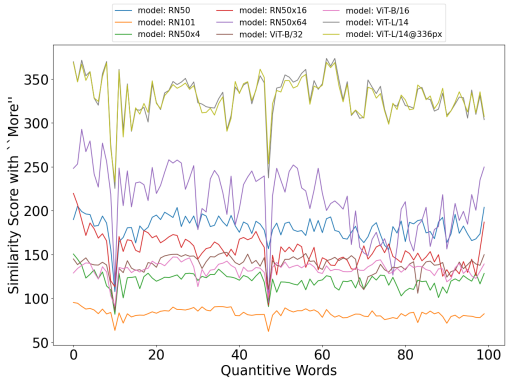

Figure 3: How “more” for selected different words.

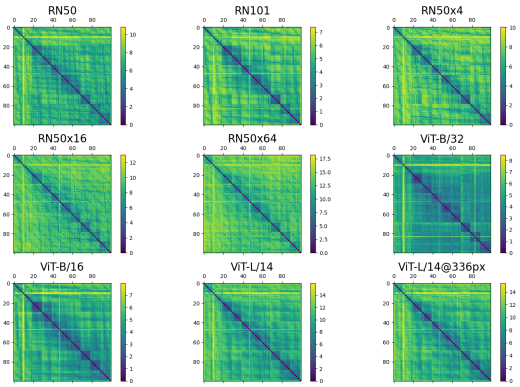

Figure 4: Distance between different quantity words.

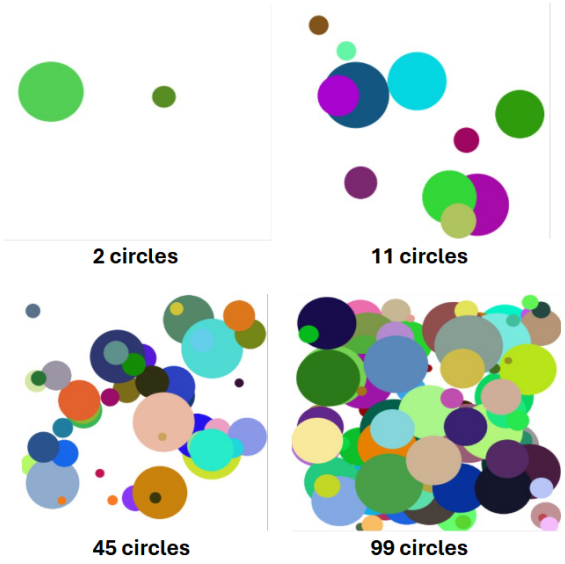

Figure 5: Examples of the images with different number of circles.

’fewer’ and ’more.’ Through this study, we aim to understand how CLIP interprets the concepts of ’few’ and ’more.’ We report the results of similarity score for the “fewer” and “more” study in Fig. 2 and Fig. 3, respectively. The statistical results reveal some interesting phenomena.

First, *CLIP cannot distinguish different quantity words well*. As the quantitative word becomes smaller or larger, the similarity score with ’fewer’ or ’more’ doesn’t decrease or increase gradually. Additionally, we can clearly see that ’zero’ demonstrates the highest similarity with both ’fewer’ and ’more.’ This could be the first evidence showing that CLIP cannot count and understand quantitative words well.

Second, *the quantity bias in the text modality is shared across different models*. We surprisingly find that the change in similarity scores for different models follows a similar trend, *i.e.*, the peaks and troughs, while differing in magnitude. For example, the values of most models at around ’fifth’ and ’thirteenth’ exhibit the maximum and minimum similarity scores, respectively, when compared with both ’fewer’ and ’more.’ These results indicate that the quantity bias is a systematic error, not a model-specific error.

To have a better understanding on the embedding of quantitive words and quantity bias, we compute the pairwise distance in $\mathcal{L}_2$ norm between different words. The results are shown in Fig. 4. With a darker color indicating a smaller distance, we can see that closely neighboring words present high similarity at the embedding level, while distant words demonstrate low similarity. On the one hand, *the high semantic similarity of closely neighboring words makes quantity comparison tricky.* On the other hand, the semantic sensitivity of the distance in text embedding causes rapid and discontinuous changes in quantity comparison. Additionally, the words around ’ten’ to ’thirteenth’ show a large distance in the embedding space compared to other words, which is consistent with previous finding in Fig. 2 and Fig. 3.

Moving to the image domain, with lacking explicit signal for quantity comparison described by pure image modality, we only compute the pair wise distance for images with different number circles. The images are randomly generated with given number of circles. We present some examples in Fig. 5.

Unlike in the text domain, as shown in Fig. 6, different images show low distance in the embedding space to each other, indicating the difficulty in differentiating them at the quantity level.

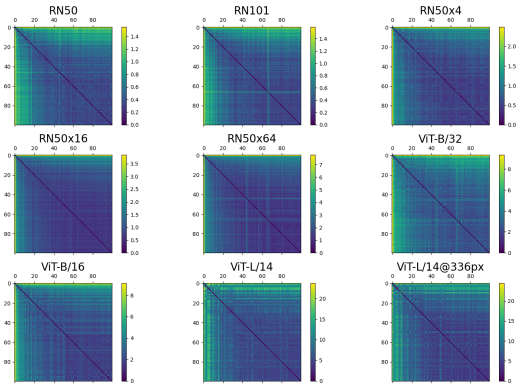

Figure 6: Distance between images with different number of objects.

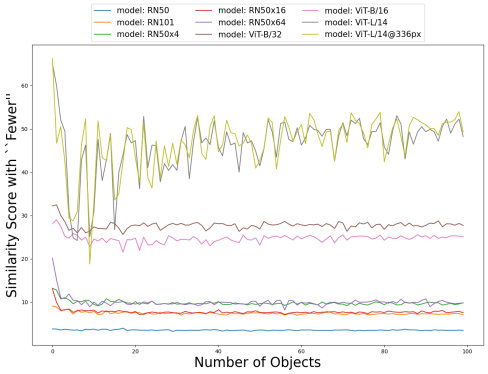

Figure 7: How "fewer" for generated objects.

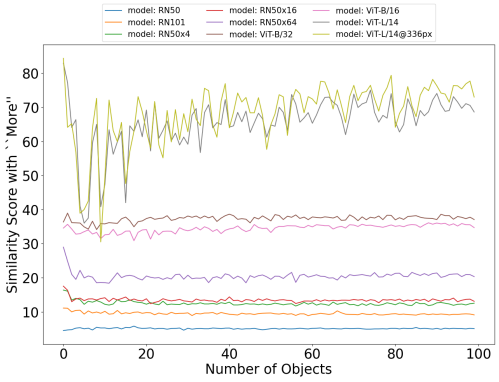

Figure 8: How "more" for generated objects.

Take-home message CLIP cannot understand the concept related to quantity in either the text or image modality, though for different reasons. In the text domain, there is high similarity between closely neighboring quantities, but a large semantic difference with distant quantity words. The irregular and noncontinuous changes between continuous quantity words make comparison difficult, leading to confusion with 'fewer' and 'more.' Conversely, in the image domain, images with different numbers of circles show high semantic similarity, making it difficult to differentiate them based on semantic differences. These factors lead to CLIP's failure in understanding quantity.

3.3 Evaluation on the multi-modal capacity

We further evaluate the quantity bias in multi-modal capacity of CLIP models. We use the quantities comparison words "fewer" and "more" to evaluate the figures with different number of circles introduced before.

We report the similarity comparison results in Fig. 7 and Fig. 8. It can be seen that most models cannot distinguish different images at the embedding level for the 'fewer' or 'more' concepts, with the similarity scores remaining smooth at a low level. Although the two ViT-L models show a large difference in word embedding between different images, they also share the same trend for the 'fewer' and 'more' concepts. This indicates that these two CLIP models learn the difference between different numbers of circles due to larger model capacity but still fail to understand quantity.

3.4 Discussion

Why this happens? We argue that two factors contribute to the ineffective learning of the quantity concept in CLIP models. First, quantity-related data are heavily limited. There are not enough data containing explicit quantity information for model learning. Many quantity words and quantity-related visual information are not considered in the model learning process, effectively making them out-of-distribution data. Second, there is a critical technical flaw in contrastive learning. While contrastive learning enables the model to learn a unified representation of different modalities, it overlooks many attributes of the inputs. For example, when contrastive learning requires CLIP to map images with circles and the text description 'circles' to the same latent space, it overlooks the comparison of the shape and number of circles in text and image modalities, leading to the existence of quantity bias in the well-trained models.

How to mitigate such bias? Based on our findings, we think there are two set of strategies for bias mitigation, including the data-centric and model-centric strategy.

- *Data-centric*: It's crucial to construct high-quality multimodal data for the training and fine-tuning of foundational models like CLIP.

Instead of simply pairing images with their names for contrastive learning, the textual descriptions should include more detailed attributes such as quantity, color, and shape. This enriched information can help distinguish the embeddings from each other in the latent space, thereby reducing embedding bias and minimizing confusion for downstream task models.

- *Model-centric*: Mitigation strategies should be tailored to specific real-world applications. For instance, addressing the counting problem highlighted in our paper, while it may be time- and computation-intensive to modify the foundational model, fine-tuning downstream task models like Stable-diffusion with carefully designed prompts, such as "many" and "fewer," can be more practical. Additionally, developers can include a regularization term that distinguishes and group different sets of quantitative words, like "one," "two" for "smaller", and "hundreds," "thousands" for "larger". This approach encourages the model to learn and differentiate quantitative concepts more effectively.

4 Conclusion

In this work, we study an interesting problem: *Can CLIP count stars?* Through extensive empirical studies on different modalities, we conclude that the CLIP models cannot understand the concept of quantity well. In the future, we will delve deeper into this quantity bias and design novel methods for efficient bias mitigation.

Limitations

CLIP is one of the most popular foundation models used in generation tasks (e.g., the development of Stable Diffusion), which motivates us to study a variety of CLIP models with different vision and text backbones in this short paper. To ensure a controlled examination of the variable in question—specifically, the number of objects in both visual and textual modalities—we employed manually constructed datasets to evaluate the quantity bias of CLIP in this preliminary research. However, real-world data presents more diversity and complexity, which were not fully captured in the simulations of this study. Future research should include more extensive results from real-world datasets and evaluate a broader range of vision-language models to provide a more comprehensive assessment.

References

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*.

Yuxuan Ding, Chunna Tian, Haoxuan Ding, and Lingqiao Liu. 2024. The clip model is secretly an image-to-prompt converter. *Advances in Neural Information Processing Systems*, 36.

Roy Ganz and Michael Elad. 2024. Clipag: Towards generator-free text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3843–3853.

Julia Guerrero-Viu, Milos Hasan, Arthur Roullier, Midhun Harikumar, Yiwei Hu, Paul Guerrero, Diego Gutierrez, Belen Masia, and Valentin Deschaintre. 2024. Texsliders: Diffusion-based texture editing in clip space. *arXiv preprint arXiv:2405.00672*.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Edgar Heinert, Matthias Rottmann, Kira Maag, and Karsten Kahl. 2024. Reducing texture bias of deep neural networks via edge enhancing diffusion. *arXiv preprint arXiv:2402.09530*.

Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. 2024. Who's in and who's out? a case study of multimodal clip-filtering in datacomp. *arXiv preprint arXiv:2405.08209*.

Peter Hönig, Stefan Thalhammer, Jean-Baptiste Weibel, Matthias Hirschmanner, and Markus Vincze. 2024. Star: Shape-focused texture agnostic representations for improved object detection and 6d pose estimation. *arXiv preprint arXiv:2402.04878*.

Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. 2018. Assessing shape bias property of convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1923–1931.

Mahyar Khayatkhoei and Ahmed Elgammal. 2022. Spatial frequency bias in convolutional generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7152–7159.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Pengkun Liu, Yikai Wang, Fuchun Sun, Jiafang Li, Hang Xiao, Hongxiang Xue, and Xinzhou Wang. 2024. Isotropic3d: Image-to-3d generation based on a single clip embedding. *arXiv preprint arXiv:2403.10395*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Oscar Michel, Anand Bhattad, Eli VanderBilt, Ranjay Krishna, Aniruddha Kembhavi, and Tanmay Gupta. 2024. Object 3dit: Language-guided 3d-aware image editing. *Advances in Neural Information Processing Systems*, 36.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

Md Mehrab Tanjim, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, and Garrison W Cottrell. 2024. Discovering and mitigating biases in clip-based image editing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2984–2993.

Zeliang Zhang, Mingqian Feng, Zhiheng Li, and Chenliang Xu. 2024a. Discover and mitigate multiple biased subgroups in image classifiers. *arXiv preprint arXiv:2403.12777*.

Zeliang Zhang, Rongyi Zhu, Wei Yao, Xiaosen Wang, and Chenliang Xu. 2024b. Bag of tricks to boost adversarial transferability. *arXiv preprint arXiv:2401.08734*.

Appendices

A More failure examples of CLIP-guided Stable-diffusion for image generation

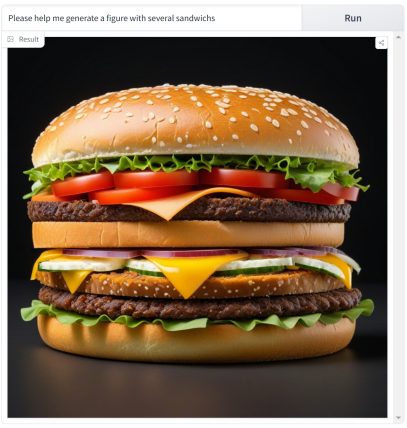

Figure A1: Prompt: Please help me generate a figure with **several** sandwiches.

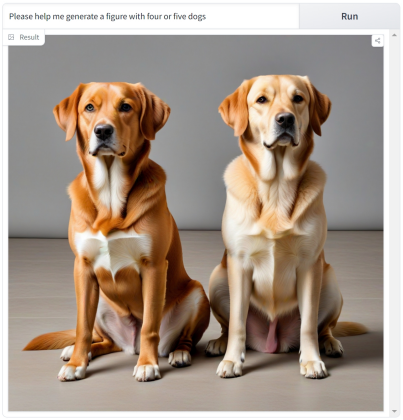

Figure A2: Prompt: Please help me generate a figure with **four** dogs.

Figure A3: Prompt: Please help me generate a figure with **ten** desks.