Improving Object Counting in Visual Langauge Models

Shreya Sivakumar, Jade Dorsainvil

October 5, 2025

Abstract

Despite there being much rapid progress in multimodal AI, it's clear that vision-language (VLMs) like LLaVA-1.5 still seem to struggle with visually grounded reasoning tasks like object counting. Failure to count properly causes difficulty in understanding the scenes, and safety. VLMs can hallucinate counting, or overlook objects. In this project, we will check whether two simple prompting types: Chain of Thought (CoT) reasoning and visually grounded reasoning. With the help of pre-existing COCO bounding boxes, we will see how to improve counting accuracy. We will compare: 1) direct answers, 2) step-by-step reasoning, and 3) grounding-assisted reasoning. We will measure absolute counting error and qualitatively analyze the resonance. The main goal is to determine if structure and grounded reasoning can help mitigate hallucinations without requiring heavy training or segmentation pipelines.

1 Problem Motivation

Our main focus for this project is to understand how the vision language models perform in generalizing outputs and gathering information from visually appealing objects and pictures. Visual language models perform consistently moderate on general visual questions but fail on reasoning performing well structured reasoning tasks such as counting, especially when scenes and objects aren't clear which include occlusion - hidden objects, clutter, or similar-looking objects. This output suggests that current VLmodels have to test with various tests to understand and improvise their word from the groundings, and therefore cannot be totally reliable for applications since they are limited in resources to reason with actual visual evidence especially. People are relying on self-driving cars, and assistive systems that require VLMs to perform appropriately. Additionally prior work provides results that models were both hallucinationing and provided inconsistent object grounding (Rohrbach et al., 2018; Liu et al., 2023), but these errors were able to be reduced due to testing and operating with various results and methods. Our motivation for this project is to test with lightweight, practical methods for improving VLM reliability without requiring complex pipelines.

2 Research Questions

Within our research, we will address the following questions:

1. Does chain-of-thought prompting actually improve the ability to object-counting accurately and compare it to direct responses?
2. Does grounding the reasoning using COCO bounding boxes help reduce hallucinations?
3. How do reasong traces differ between direct, CoT, and grounded methods? What type of scenses or even object will cause LLaVa-1.5 to fail?

3 Proposed Solution

Our methodology is based of three main strategies, and based on these we have evaluated our results using LLaVA-1.5 model:

Baseline. We using the model to answer our question: We ask: “How many object are in this image?” The object on the image was apple The model provided a final count of apples/objects that is on the image.

Chain-of-Thought Reasoning. Secondly, we initiated the model to detect new objects themselves to understand how they look on the objects and based on what appearance. We asked to list the objects as you see and describe where they are located and how many objects of each variety.

Grounded Reasoning with COCO Bounding Boxes. We implemented a JSON file that stores the images and information and leverages existing COCO annotations to the model. From the images we prompted we asked the model to count the objects that were highlighted in the image and by referencing only these boxes. This forces the model to align its reasoning with predefined visual evidence that was given to test without needing to go through and run any additional vision models. Dataset and Evaluation

To evaluate our final output to understand its reliability and capability we used 10–15 images from the MS COCO 2017 dataset. And for each image, we test all the three methodology conditions and compute: Absolute counting error Qualitative reasoning accuracy such as hallucinated instances, incorrect references

4 Relation to Prior Work

In Zhang research, it was noticed that VLMs tend to hallucinate certain objects or even miscount instances when they are in cluttered scenes (Zhang et al.,2023). With the idea of chain-of-thought, it has certain techniques that can improve reasoning though in more text task (Wei et al., 2022), but visual reasoning remains a bit unreliable (Gao et al., 2023). Meanwhile, its noteworthy to see how visual grounding and segmentation systems like DINO (Zhang et al., 2022), SAM (Kirillov et al., 2023), and region-guided pipelines (PseCo; Xu et al., 2022) demonstrates the usefulness of explicit region supervision. In our project, we had drawn on much of these inspirations and research from these works, but we focused on more lightweight, prompt-only versions.

5 Blocks and Anticipated Challenges

1. LLaVA sometimes generates overly long or unfocused reasoning steps when using chain-of-thought style prompts, which may make it harder to evaluate responses consistently.

2. The bounding boxes we generate may not cleanly capture the target objects, especially in cases involving occlusion or cluttered scenes, which can introduce noise into the model’s inputs.

3. Even with grounded or region-specific prompts, the model may still infer or hallucinate details that are not actually present within the selected visual region.

4. Maintaining a consistent experimental setup across different prompt types and model runs requires careful control of variables, since small changes can affect output behavior.

5. Due to project time constraints, the dataset must remain relatively small, which limits the breadth of scenarios we can test and may impact the generalizability of our results.

6 Initial Ethical Considerations

When it comes to how important the idea of counting in real-world implications, we see it in surveillance, crowd monitoring and even policing. An error can really lead to misinterpretations or even overconfidence in a models outputs. Even using COCO bounding boxes has to look at images of people, which can be seen as a privacy concern.

References

[1] T. Chen, S. Kornblith, M. Norouzi, & G. Hinton (2020). *A Simple Framework for Contrastive Learning of Visual Representations*. arXiv:2002.05709. https://arxiv.org/abs/2002.05709

[2] K. He, H. Fan, Y. Wu, S. Xie, & R. Girshick (2020). *Momentum Contrast for Unsupervised Visual Representation Learning*. CVPR. https://arxiv.org/abs/1911.05722

[3] J.-B. Grill, F. Strub, F. Altché, et al. (2020). *Bootstrap Your Own Latent (BYOL)*. NeurIPS. https://arxiv.org/abs/2006.07733

[4] Y. Tian, X. Chen, & S. Ganguli (2020). *What Makes for Good Views for Contrastive Learning*. NeurIPS. https://arxiv.org/abs/2005.10243

[5] P. Khosla, P. Teterwak, C. Wang, et al. (2020). *Supervised Contrastive Learning*. NeurIPS. https://arxiv.org/abs/2004.11362

[6] X. Zhang & M. Maire (2020). *Self-Supervised Neural Networks: A Survey*. arXiv:2010.05113. https://arxiv.org/abs/2010.05113

[7] A. van den Oord, Y. Li, & O. Vinyals (2018). *Representation Learning with Contrastive Predictive Coding*. arXiv:1807.03748. https://arxiv.org/abs/1807.03748

[8] Hzzone (2024). *PseCo: Pseudo-Contrastive Learning for Object Representation*. GitHub repository. https://github.com/Hzzone/PseCo