VisionClue: Improving Multi-modal Language Model on Object Counting with Self-Generated Hints

Yinuo Zhao
University of Toronto

Yuanyi Liu
University of Toronto

Abstract

VisionClue is a two-stage prompting strategy that can improve the performance of multi-modal language model on object counting from images without needing more data. In the first stage, the model sees the image and generates description and hints, in the second stage, the model receives the image and the previously generated information to provide an answer. We found direct strategical hints to be most helpful and incorporating it decreased the root mean squared error by 30%.

1 Introduction

Counting the number of objects in images is a task with many practical applications, including but not limited to crowd management, traffic analysis, and ecological monitoring. Humans excel at this task due to our ability to "think," which involves extracting and utilizing various types of information from both the image and auxiliary sources and adopting the most suitable strategy to carry out complex reasoning. While this may seem easy, it is actually challenging for many state-of-the-art machine intelligence systems.

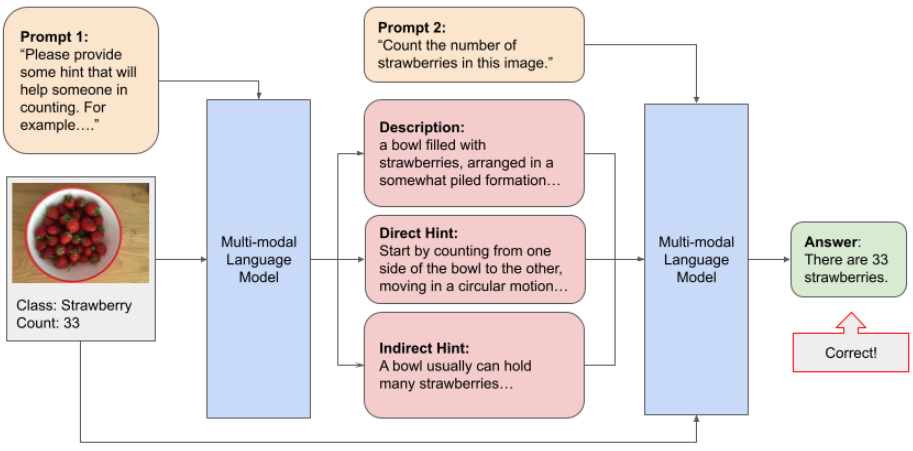


Figure 1: VisionClue uses the model to generate hints, then passes the hint and the image to the model.

The presence of multi-modal language models has made it possible to combine visual perception and linguistic interpretation, which are the two essential parts that object counting requires. However, the traditional ways of simply passing the image to the model and asking it to count the number of

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

objects inside undermine the potential of the text interpretation ability of the multi-modal language model, resulting in suboptimal performance. To address this, we present VisionClue, a novel two-stage approach that enhances the accuracy of counting using self-generated side information. In the first stage, the image is passed to the model with instructions to create three types of information: description, direct hint, and indirect hint. In the next stage, the generated information is passed together with the image to the model again to obtain a final answer. The complete workflow is shown in Figure 1. In this study, we systematically explored how different types of information can boost accuracy. The results are compared against the vanilla model performance, where no extra information is given, and with the human performance benchmark.

VisionClue is a highly effective approach that enhances performance without additional data collection and is applicable to various object types. This method is particularly advantageous in scenarios where only image data is available since it utilizes self-generated rather than externally collected information. Moreover, its robust generalization capabilities eliminate the necessity to train or fine-tune specific models for different object categories, significantly expanding its application scope. All related data and code are accessible at here.

2 Related Work

Many supervised learning frameworks have been designed for object counting, such as generating a density map across an image using a newly developed loss function [2]. Although such methods are efficient and adaptable, they cannot utilize textual data. Vision Transformers (ViTs) have demonstrated significant advancements over traditional CNNs, showing superior performance on various image recognition benchmarks, including ImageNet, while requiring fewer resources for training [1].

Following these developments, multi-modal language models have been introduced, merging the strengths of LLMs in processing text with the capabilities of ViTs in image handling. Notable examples include OpenAI's GPT-4, DeepMind's Flamingo, and Google's PaLM-E, which have gained popularity for their ability to process and generate content across diverse data types. This integration facilitates more robust and contextually aware interactions, enhancing applications in visual estimation. However, there is still a gap in effectively utilizing the textual interpretation capabilities of these models for object counting. VisionClue addresses this gap, serving as an analogy to a series of intermediate reasoning steps akin to the Chain-of-Thought prompting strategy, which has been shown to significantly improve the complex reasoning abilities of LLMs [6].

3 Methodology

The dataset we used for this study is adapted from FSC147 [4], which was originally compiled to be a Few-Shot counting dataset. It contains 6,135 images that belongs to 147 different categories. Each image in the dataset is annotated with bounding boxes for exemplars and a dot-annotation marking each object instance. For the purpose of our study, we randomly subset 300 images and extracted the true object counts from the annotations.

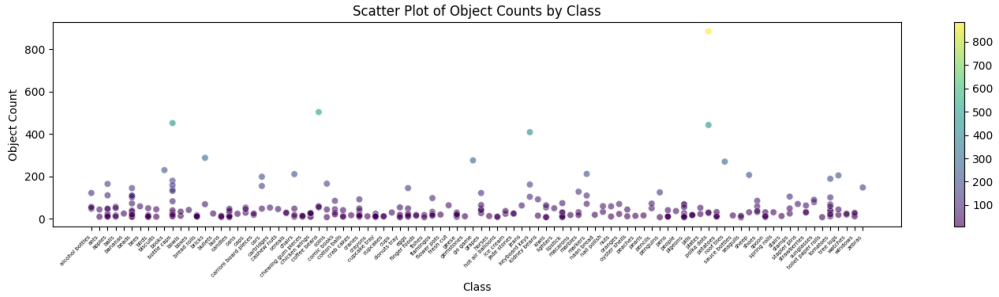


Figure 2: Between object categories, the range of object counts differs.

The truncated dataset has 96 unique categories with object counts ranging from 7 to 885, averaging at 52. Within each category, the object counts vary as shown in Figure 2, and the range differ between

categories. While the dataset has a wide range of object counts, most of the images have low counts with few above 200. A better illustration can be found at Figure 5.

Given the increased difficulty in object counting as numbers rise, we divided the images into three groups based on object count to ensure a more accurate performance analysis. These groups are: images with fewer than 20 objects, images with 20 to 100 objects, and images with more than 100 objects. Table 1 summarizes the distribution of images within each group.

	Group 1	Group 2	Group 3
Object Count	$[0, 20)$	$[20, 100)$	$[100, \infty)$
Number of Images	122	139	39

Table 1: The 300 images are divided into 3 groups by object count.

The multi-modal language model we used is GPT-4o mini[3]. We opted not to specify further parameters to avoid imposing restrictions that prevent us from evaluating the performance of a vanilla model.

Recall from Figure 1 that two prompts were employed in VisionClue. The first prompt adapted a One-Shot design with a knowledge prompting mechanism[5]. It asks the model to generated 3 types of information: description, direct strategical hint, and indirect contextual hint. The description is a detailed explanation of the object in the image, including position, orientation, color and other characteristics. Direct strategic hints are strategies to count the objects, such as ordering, grouping and recognizing patterns. Indirect contextual hints are domain specific background knowledge related to the object, such as typical size and behavior of occurrence. This prompt includes definitions of these three information types along with a full text-only example. The second prompt merges the provided information and prompts the model to count the number of objects in the image. Importantly, the category of the objects is specified in both prompts. The complete templates for these prompts are included in the Appendix.

We conducted five experiments to gather responses from the vanilla model as the baseline, and from models given all three types of information, only the description, only the direct hint, and only the indirect hint. We employed the root mean squared error (RMSE) as the primary metric to evaluate the effectiveness of each approach due to its sensitivity to larger errors. When the predicted count is empty, we ignore that entry from the error calculation.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

Additionally, we tracked the number of images the model failed to count, as we are interested in whether providing extra information helps the model to accomplish tasks it otherwise could not. To provide a clear understanding of the challenge of object counting within these 300 images, we also measured human performance as a benchmark.

4 Results

We observed that GPT-4 tends to underestimate the number of objects in images, regardless of the type of information provided. This tendency is illustrated in Figures 3 and 4, where the majority of data points fall below the diagonal line, indicating a perfect count.

The vanilla GPT-4, without any additional hints, achieved an RMSE of 66.82 across all images, which is double the human error of 34.90. It also exhibited a trend of increasing relative performance with higher true counts: its performance on small count images (0 to 20) was approximately four times worse than that of humans. For images with 20 to 100 objects, the error was triple that of humans, and for those with more than 100 objects, the error was twice as high. The specific RMSE values are detailed in the first two rows of Table 2. Additionally, it is important to note that while humans provided responses for all images, often estimating rather than precisely counting, the vanilla GPT-4 was unable to make such estimations and failed to respond to 14 images.

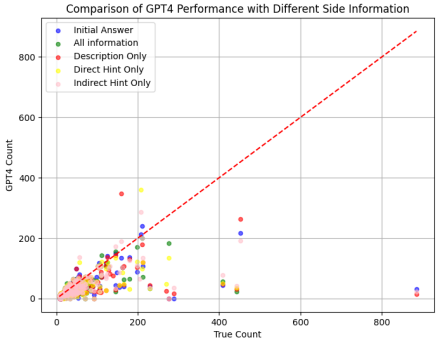


Figure 3: GPT4 counts vs true counts on all images.

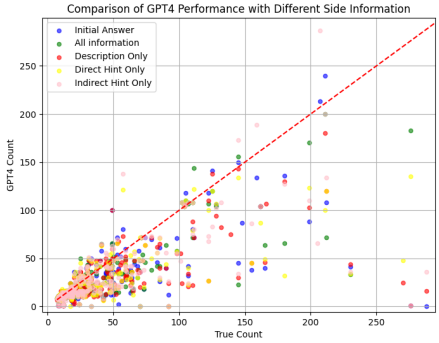


Figure 4: GPT4 counts vs true counts, zoomed on smaller count images.

Method	All	$[0, 20)$	$[20, 100)$	$[100, \infty)$	Failed to Count
Human	34.90	0.54	7.41	95.78	0
GPT4: vanilla	66.82	4.134	23.90	203.23	14
GPT4: description only	72.80	3.98	22.25	209.21	6
GPT4: direct hint only	46.36	3.51	22.73	142.40	23
GPT4: indirect hint only	69.91	3.83	22.57	204.78	11
GPT4: all information	42.08	3.48	20.80	134.03	36

Table 2: Root mean squared error of the generated counts.

When provided only with descriptions, GPT-4's overall performance worsened to an RMSE of 72.80. It exhibited a higher error (209.21) on images with more than 100 objects, although the difference was marginal. For images with fewer than 100 objects, it showed a slight improvement over the vanilla GPT-4. This model configuration failed to count just 6 images. When provided only with direct hints, the overall error significantly dropped by 30% from the vanilla model. This improvement was largely due to increased accuracy on images with more than 100 objects, where the error decreased from 203.23 to 142.40. However, this approach resulted in the model failing to count on 23 images. When given only indirect hints, the model performed similarly to the vanilla model but with slightly higher overall error and fewer failed counts. Lastly, when prompted with all three types of information, GPT-4 demonstrated the best performance of all the experiments across all object count ranges. However, the improvement was not significantly different from using direct hints only. The overall RMSE was 42.08, but this configuration resulted in the highest number of failures at 36 images.

5 Discussion

Our experiments demonstrate that VisionClue can enhance the performance of multi-modal language models, particularly GPT-4, in object counting through the use of self-generated side information. This method involves two prompts: the first to gather information and the second to combine this information with the image to aid in counting. We found that direct strategical hints were the most beneficial, reducing the overall RMSE of the dataset by 30%.

Additionally, we observed a trade-off between RMSE and the number of images the model failed to count. The results suggest that achieving a smaller error often comes at the expense of having more uncounted images. This pattern raises a reasonable concern that the model might simply choose not to answer for all images that are challenging to count, thereby achieving a lower error rate but not necessarily delivering desirable outcomes. This study is constrained by the limited characterization and measurement of images in the dataset, as well as thorough understanding of the reasoning logic behind the model. Enhancing these aspects could reveal deeper insights into which types of images are considered "hard to count" and why GPT-4 fails to count them.

Despite VisionClue's effectiveness in improving accuracy, its performance still does not approach that of humans. However, this research points to a promising direction for further investigation in this area.

References

[1] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[2] Victor Lempitsky and Andrew Zisserman. "Learning To Count Objects in Images". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty et al. Vol. 23. Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper_files/paper/2010/file/fe73f687e5bc5280214e0486b273a5f9-Paper.pdf.

[3] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[4] Viresh Ranjan et al. "Learning To Count Everything". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[5] Yuheng Shi, Xinxiao Wu, and Hanxi Lin. *Knowledge Prompting for Few-shot Action Recognition*. 2022. arXiv: 2211.12030 [cs.CV]. URL: https://arxiv.org/abs/2211.12030.

[6] Jason Wei et al. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL]. URL: https://arxiv.org/abs/2201.11903.

Contribution

- Project Idea and Study Design: Yinuo Zhao, Yuanyi Liu
- Dataset: Yinuo Zhao
- GPT4 experiments: Yinuo Zhao
- Performance Evaluation: Yuanyi Liu
- Report: Yinuo Zhao (Methods and Results), Yuanyi Liu (Introduction and Literature Review)

Appendix

A. Dataset

Figure 5 is a box plot demonstrating the distribution of object counts in the dataset we used.

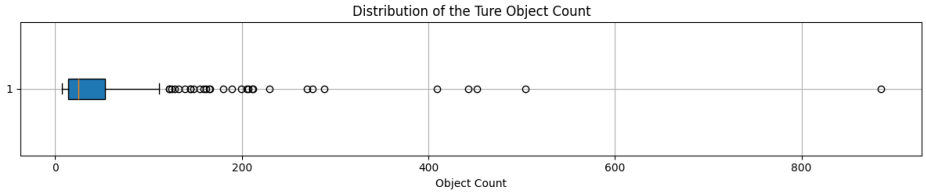


Figure 5: The distribution of object count is very right skewed.

B. Prompts

Figure 6 contains the templates for the two prompts used in VisionClue. Note that when no extra information is given, Prompt 2 only has the last sentence.

Please generate informations that can help someone on counting the number of {object_name} in this image. You need to provide the following:

1. Description: details of the objects in this image.
2. Direct hint: guidelines on methods to count the number of objects
3. Indirect hint: the contextual or background information about the object that will help in counting.

For example, if you are seeing an image of geese, you should provide the following:

1. Description: The image features a group of Canada geese in flight against a clear blue sky. The geese are dispersed across the image in various flight positions, with their wings in different phases of the flapping cycle.
2. Direct hint: To count the number of geese, start from one corner of the image and move your eyes in a grid-like pattern, left to right, top to bottom, marking each bird as counted to avoid recounting the same goose.
3. Indirect hint: Geese often travel in V-shaped formations or smaller groups, which can help you estimate their numbers more effectively. When counting, keep in mind that the number will likely reflect typical group sizes seen in nature, rather than a sparse or overly dense arrangement.

You have the following information available to help you:

1. Description (if given)
2. Direct hint (if given)
3. Indirect hint (if given)

Please count the number of {object_name} visible in this image and respond with only the numeric answer.

Figure 6: Template of Prompt 1 at the top, template of Prompt 2 at the bottom

C. Generated Information by GPT4

We provide two examples of the information generated by GPT4:

Figure 7: Image A

Figure 8: Image B

Image	Description	Direct Hint	Indirect Hint
A	The image features an egg carton containing 18 eggs arranged in a grid pattern. There are both brown and white eggs, displayed in a manner that allows for easy visual identification. The carton is made of cardboard with segmented sections to hold each egg securely. The background shows a wooden table surface.	To count the number of eggs, start from one side of the carton and methodically count from left to right in each row, moving down to the next row until you reach the bottom. Keeping a finger or a marker at the last counted egg can help avoid recounting.	Egg cartons typically come in standard sizes, commonly holding 12, 18, or 24 eggs. This can provide context to your counting process, as you can expect the arrangement to fit within these common quantities. Additionally, maintaining focus on distinguishing between the brown and white eggs could aid in managing the count more effectively.
B	The image showcases a rustic wooden surface with several almonds, both shelled and unshelled, scattered around. A small glass bottle containing yellow oil, presumably almond oil, is positioned to one side. The almonds vary in appearance; some are whole while others have their shells cracked open, revealing the inner nut.	To count the almonds in the image, visually separate the nuts from the background. Focus on one type of nut at a time. You may want to group them based on whether they are shelled or unshelled. Count systematically from left to right or top to bottom to ensure each nut is counted only once.	Almonds typically grow in clusters, and a common serving size may help gauge the quantity. Considering that almonds can be served whole or in recipes, think about how they are usually packaged or sold (e.g., in bags of 100 grams) to estimate the number you see in the image based on their typical serving sizes.

D. Performance Analysis

Figure 9 and Figure 10 are scatter plots demonstrating human performance on object counting versus the true count. Plot on the left-hand side is complete and shows all 300 images. Plots on the right-hand side is zoomed on lower-count images for better visualization.

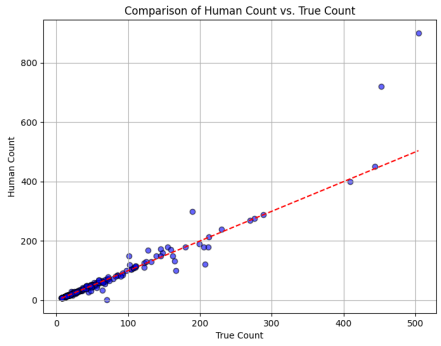


Figure 9: Human vs true: all images.

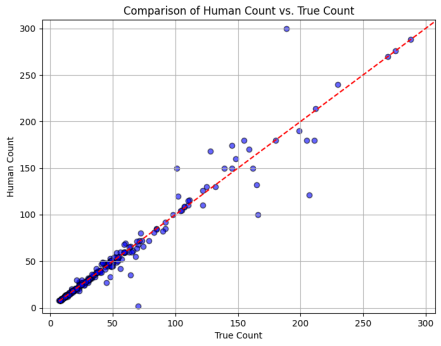
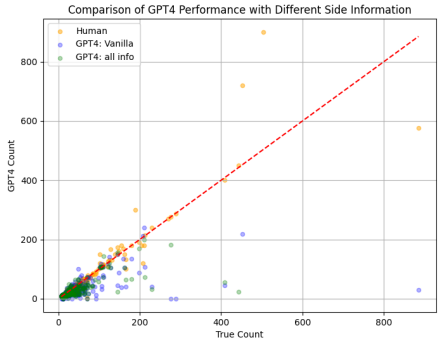
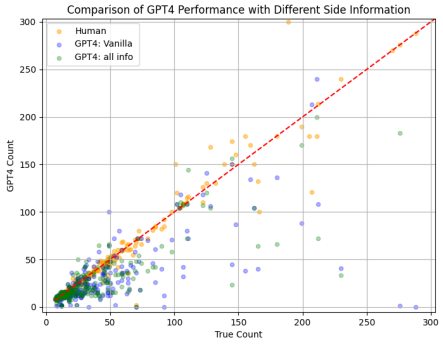
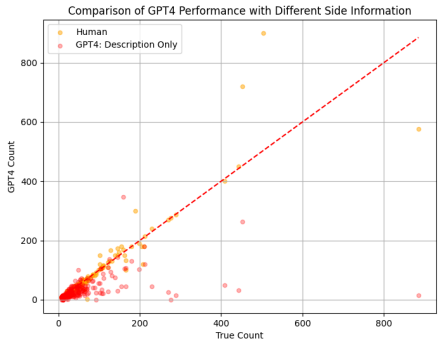


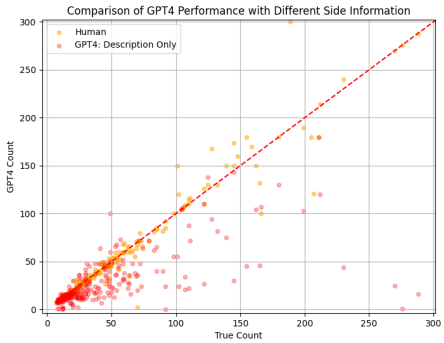
Figure 10: Human vs true: low count images

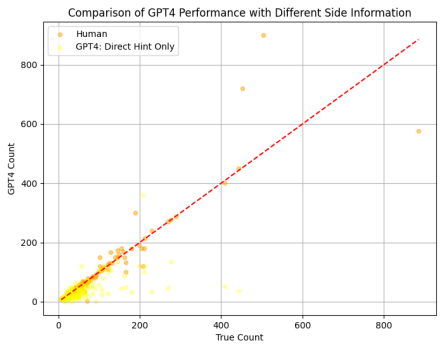
Below are Scatter plots demonstrating the performance of GPT4 with various types of information provided. Right side plots are the same as left side except zoomed on low count images as before.

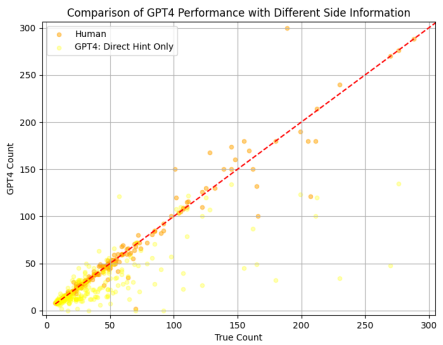

Comparison of GPT4 Performance with Different Side Information
Human
GPT4: Vanilla
GPT4: all info
GPT4 Count
True Count

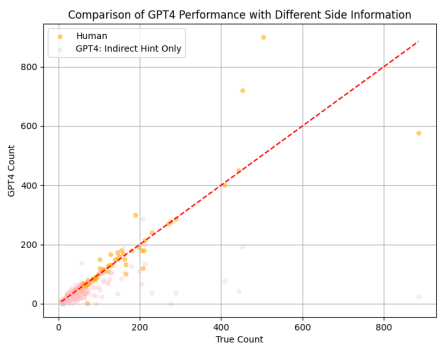

Comparison of GPT4 Performance with Different Side Information
Human
GPT4: Vanilla
GPT4: all info
GPT4 Count
True Count

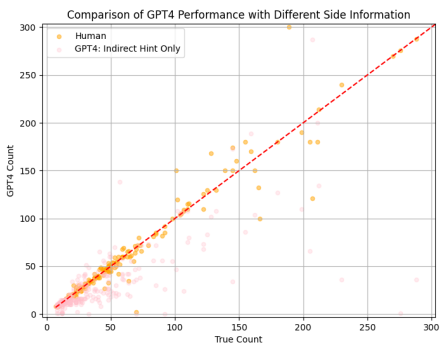

Comparison of GPT4 Performance with Different Side Information
Human
GPT4: Description Only
GPT4 Count
True Count


Comparison of GPT4 Performance with Different Side Information
Human
GPT4: Description Only
GPT4 Count
True Count


Comparison of GPT4 Performance with Different Side Information
Human
GPT4: Direct Hint Only
GPT4 Count
True Count


Comparison of GPT4 Performance with Different Side Information
Human
GPT4: Direct Hint Only
GPT4 Count
True Count


Comparison of GPT4 Performance with Different Side Information
Human
GPT4: Indirect Hint Only
GPT4 Count
True Count


Comparison of GPT4 Performance with Different Side Information
Human
GPT4: Indirect Hint Only
GPT4 Count
True Count