Spread them, Count Them: Class-agnostic Fine-grained Visual Counting

Viet Bach Tran
University of Adelaide

Gia Khanh Nguyen
University of Adelaide

Minh Hoai Nguyen
University of Adelaide

Abstract

Visual counting presents a significant challenge in computer vision, particularly in scenarios where similar categories of objects coexist, leading to difficulties in fine-grained detection within the same class or among visually similar objects. While existing methodologies demonstrate the ability to count previously unseen objects, they often fall short in distinguishing subtle differences that are critical for accurate object identification. This paper aims to enhance the fine-grained capabilities of object counters in class-agnostic contexts by introducing FGC200K, a comprehensive dataset designed specifically for the development and training of fine-grained object counting models. Comprising over 200,000 images featuring 100 pairs of visually similar categories, FGC200K provides a diverse and complex resource that enables the advancement of more accurate models, ultimately improving decision-making in applications such as inventory tracking, quality control, and crowd management. Code and dataset is available at https://github.com/Theskrtnerd/fine-grained-visual-counting

1. Introduction

Visual object counting is a core problem in computer vision, with applications in crowd management, inventory tracking, and environmental monitoring. While progress has been made in agnostic counting—where models can count objects from various categories without class-specific training—fine-grained counting remains a challenge. Specifically, distinguishing between visually similar objects with subtle differences is difficult, especially without visual exemplars, posing challenges in real-world scenarios where nuanced identification is essential.

Existing datasets for object counting lack the diversity and complexity needed to address these fine-grained cases. Models trained on them struggle to differentiate between similar objects—such as two types of tools or similarly shaped products—leading to reduced accuracy in practical applications.

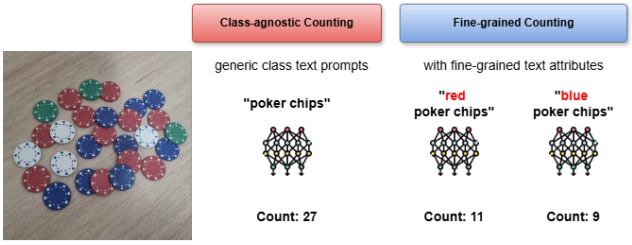

Figure 1. Example of coarse-grained vs. fine-grained counting: The fine-grained task focuses on counting only "red poker chips", adding attribute-based constraints for improved real-world applicability.

To address this, we introduce a novel dataset FGC200K with over 200,000 images, featuring 100 pairs of visually similar categories. Each pair highlights subtle differences in shape, size, texture, or color, reflecting real-world complexities. The dataset spans diverse environments with varying backgrounds and occlusions, providing a robust foundation for training models to perform precise object differentiation and counting.

Our dataset bridges the gap between general counting and fine-grained identification, offering a valuable resource for advancing research in object counting. By capturing nuanced distinctions, it supports the development of models capable of attribute-based counting, improving decision-making in applications that rely on precise, detailed recognition.

As illustrated in Figure 1, fine-grained counting focuses on distinguishing and counting objects based on specific attributes, such as color, size, or texture, which are often overlooked in general counting tasks. For example, instead of counting all "poker chips" in a scene, fine-grained counting targets only "red poker chips," even in scenarios where multiple colors are present. This added layer of complexity aligns with real-world challenges where precise differentiation is essential.

In addition to attribute-based counting, the FGC200K dataset also addresses the challenge of distinguishing between different objects that are visually similar. Many real-world tasks involve separating objects with overlapping vi-

sual features—such as distinguishing between two types of screws or tools. Our dataset includes 100 pairs of visually similar categories to reflect these complexities, ensuring models are capable of both fine-grained identification and cross-category differentiation. This design enables models trained on FGC200K to achieve higher accuracy in practical applications, where both nuanced distinctions and visual similarities between objects play a critical role in decision-making.

2. Related Work

2.1. Class-agnostic Object Counting

Traditional object counting methods have focused on class-specific counting, where models are trained to count objects belonging to specific categories (e.g., cars, people). However, the focus has shifted towards the more challenging task of class-agnostic object counting [1, 8, 11, 15].

Class-agnostic object counting aims to develop models that can predict the number of objects from any category, including those not encountered during training. This approach is inherently more complex as it requires models to generalize across a wide range of object types. As a result, large and diverse datasets, such as FSC-147 [11], play a crucial role in effectively training these models.

Recent techniques, such as ZSC [15] and CountGD [1], utilize these diverse datasets to build robust models capable of counting objects outside their training distribution. These methods have shown promising results, enabling models to handle novel object types and generalize effectively to real-world scenarios.

However, despite their importance, the existing datasets remain limited in size and diversity, constraining the generalization capabilities of current models. Furthermore, the repeated use and recycling of the same datasets across different studies highlight the need for new, larger, and more varied datasets to push the boundaries of class-agnostic object counting and enable more robust performance in real-world applications.

2.2. Text-based Counting Methods

With the advancement of Vision-Language Models (VLMs) like CLIP and GroundingDINO [10], text-based counting methods have become a promising approach for object counting. These methods leverage textual prompts to interact with image features, enabling zero-shot counting and enhancing class-agnostic object counting capabilities [2, 7, 13, 15].

One of the strengths of text-based methods is their flexibility in using various prompts to differentiate objects within the same image. For example, CLIP-based models have been fine-tuned with contrastive loss between true and counterfactual prompts alongside the standard text-image contrastive objective, enabling them to count up to ten objects. Methods like ZSC [15], CounTX [2], and CLIP-Count [7] embed class names or text descriptions through CLIP's text encoder to produce density maps, allowing for zero-shot counting across different object categories.

Building on this foundation, we tackled the problem by creating a fine-grained dataset that incorporates attributes such as shape, color, and texture to fully leverage the strengths of text prompts. This approach allows for more precise identification and counting, even for visually similar objects. By focusing on fine-grained differentiation, our method stands apart from prior approaches and improves performance in complex real-world scenarios.

2.3. Fine-grained Counting Tasks

The Fine-grained Counting Task is still a relatively new problem and has so far been addressed primarily by GroundingREC [4], which introduced the REC-8K dataset. REC-8K serves as a benchmark for referring expression counting (REC), where the task involves not only counting objects but also identifying specific instances based on descriptive attributes. This extension makes the task more challenging by requiring models to distinguish between objects with subtle variations in features like shape, size, and color.

REC-8K was constructed by selecting and re-annotating images from existing datasets such as FSC-147 [11], CARPK [6], JHU Crowd [12], and NWPU [14]. This reuse of images demonstrates the limitations in diversity and novelty, as the dataset relies heavily on previously explored data sources. While it introduces fine-grained annotations and attributes, REC-8K remains constrained by the size and diversity of the original datasets it draws from. As a result, there is still a need for more comprehensive datasets that capture a wider variety of real-world scenarios to improve model generalization.

Our work aims to address this gap by developing a novel dataset with greater size and diversity, specifically tailored for fine-grained object counting. Unlike REC-8K, which reuses images from previous datasets, our dataset introduces new real-world scenarios across domains such as industrial environments, outdoor events, and retail spaces. This expanded scope will enable models to better generalize across unseen conditions and enhance performance in real-world applications. By incorporating new domains and increasing diversity, our dataset pushes the boundaries of fine-grained object counting, offering a more robust benchmark for future research.

3. FGC200K Dataset

3.1. Motivation

Figure 2. Example images from FGC200K illustrating various attribute distinctions within the same class.

Class-agnostic visual counting has been explored using various datasets, with FSC-147 [11] being one of the most commonly used. However, while FSC-147 has proven effective for class-agnostic counting tasks, it falls short in addressing fine-grained counting. Similarly, , widely employed in visual counting studies, does not adequately control for intra-class similarity. This limitation impacts its ability to capture subtle variations between visually similar objects. The REC-8K [3] dataset, specifically designed for fine-grained counting, introduces annotations for different object attributes but still faces significant challenges. Notably, REC-8K's relatively small size (8,011 images) and its reuse of images from other datasets, such as FSC-147 and CARPK, highlight its limited scope. The lack of original data further restricts its effectiveness for comprehensive fine-grained counting tasks.

To address these challenges, we developed the FGC200K dataset, which consists of more than 200,000 images of 100 categories, by conducting our own data collection. One of the primary limitations of existing datasets is the lack of fine-grained distinctions, especially between two visually similar objects, which constrains their utility for nuanced counting tasks. By capturing the data ourselves, we introduce natural imperfections—such as varying lighting, background clutter, and image noise—that better reflect real-world conditions. Furthermore, we selected object categories that are both practical and relevant to everyday scenarios, including food, toys, and household items. Although we did not include every possible object type (e.g., people or animals), the dataset covers a broad range of commonly encountered categories, enhancing its applicability in real-world contexts (Figure 2).

Controlling the entire data collection process not only provides flexibility for future research but also enables the extension of the dataset and experimentation with new algorithms. Our design focuses on practical use cases, such as monitoring game pieces before and after play, accounting for rearrangements, and managing noise. This approach allows the dataset to be used beyond counting, supporting tasks like verification and other downstream applications, thereby broadening the scope of visual counting research.

3.2. Data Generation Process

Creating 200,000 individual images manually would have been impractical, given the time and effort required to set up scenes, track labels, and capture photos accurately. To overcome these challenges, we adopted a combinatorial approach (as illustrated in Figure 3) to generate a diverse and scalable dataset efficiently. Specifically, we identified 100 pairs of visually similar objects. Each pair was designed to reflect subtle visual differences, ensuring relevance to fine-grained counting tasks. For every object pair, we created 10 distinct label sets, representing different object counts of each pair to simulate a variety of counting scenarios. This labeling scheme was essential for capturing nuanced variations in object quantities that existing datasets often overlook.

For every unique label set, we generated two distinct

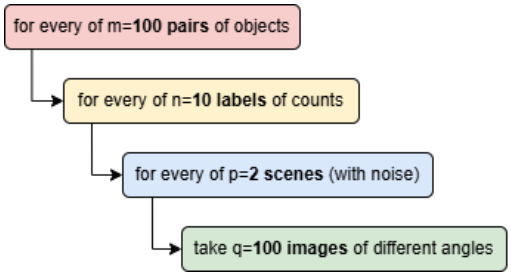

Figure 3. Overview of the data collection process

scenes by rearranging the objects, introducing noise (such as random misplacements or occlusions), and varying the background settings. This scene variation was intended to simulate real-world complexities, such as cluttered environments and changing conditions.

Rather than capturing individual static images, we recorded video clips for each scene. These videos were then split into 100 frames per clip, which allowed us to efficiently produce a large number of labeled images while maintaining variability. This method introduced natural variations in object positioning, lighting conditions, and camera angles that would have been difficult to replicate through manually staged photographs.

Our video-based strategy not only ensured scalability but also enhanced the dataset's diversity. By leveraging subtle changes in consecutive frames, the resulting images exhibit rich variability, closely resembling real-world scenarios. This variability makes the dataset well-suited for practical applications, where objects are often encountered in dynamic or imperfect conditions. Additionally, this approach provides an efficient framework for generating large-scale labeled datasets, setting a foundation for future research in fine-grained visual counting tasks.

3.3. Object Pair Selection

We selected 50 object categories from everyday life, such as food, toys, household items, and office supplies, to ensure diversity. These categories were used to create 100 pairs of visually similar objects, divided into two key types (as shown in Figure 4):

- **Intra-class Pairs:** Pairs of objects from the same category with subtle differences, such as color, texture, or shape. Examples include red vs. blue poker chips and spiral vs. penne pasta. These pairs challenge the model to detect fine-grained distinctions.
- **Inter-class Pairs:** Pairs of objects from different categories that look visually similar, such as coins vs. buttons or lemons vs. eggs. These pairs test the model's ability to distinguish between different objects with similar shapes or appearances.

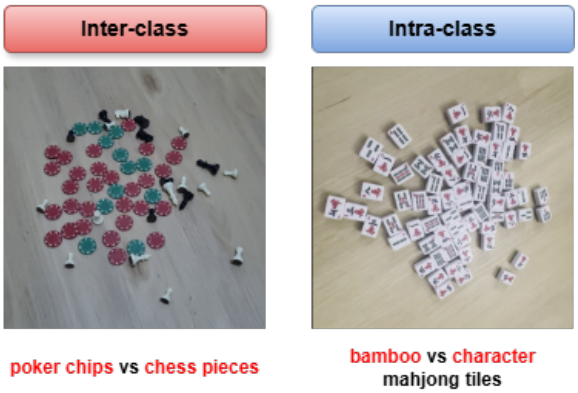

Figure 4. Examples of intra-class and inter-class pairs: Inter-class pairs involve different text prompts targeting distinct objects, while intra-class pairs focus on the same object with variations in fine-grained attributes.

For each type, we decided to find 50 pairs, making a total of 100 pairs of objects. While intra-class pairs were relatively easy to generate by identifying each object's fine-grained attributes, finding suitable inter-class pairs was more challenging due to the problem of evaluating visually similar pairs in a systematic and unbiased way.

Inter-class Pair Similarity Evaluation. To assess the similarity between objects from different classes, we utilized the ResNet-50 [5] model to extract feature representations from the images. Each image was first resized to 224x224 pixels to maintain consistency with the input requirements of the model. Afterward, the resized images were passed through the pre-trained ResNet-50 network, where the output from the penultimate layer was extracted as a 2048-dimensional feature vector for each object. These feature vectors capture essential visual characteristics and serve as concise descriptors of the images.

Using these vectors, we generated a similarity matrix that evaluates pairwise similarities between objects across different classes. The matrix enables a systematic comparison between every possible inter-class pair. For this task, we utilized cosine similarity to compute the similarity score between feature vectors. Cosine similarity is particularly effective in identifying the alignment between vectors in high-dimensional space, which ensures that the comparison is robust to potential variations in scale or brightness across images.

Each element in the similarity matrix corresponds to a score that reflects the similarity between two objects, with higher scores indicating a greater degree of similarity. This matrix not only allows us to quantify how similar objects are but also reveals relationships between distinct classes, which may highlight patterns such as overlapping visual characteristics or unexpected commonalities.

By leveraging the ResNet-50 model's feature extraction

and cosine similarity for pairwise comparison, our approach ensures a reliable evaluation of inter-class similarities. This analysis is valuable in contexts such as clustering, anomaly detection, or fine-grained classification, where understanding subtle inter-class relationships can provide critical insights. The generated similarity matrix offers a detailed view of the data, enabling further exploration of connections between objects that might otherwise go unnoticed.

3.4. Data Splits

A key challenge in partitioning the dataset was managing the inter-class relationships between visually similar objects. If objects with high similarity were split across different subsets, the model could overfit by learning class-based patterns instead of focusing on meaningful visual features. To prevent this, we ensured that visually similar objects were grouped within the same clusters. This approach promotes learning based on shared visual characteristics rather than relying on class-specific attributes.

To organize the dataset effectively, we utilized a similarity matrix generated from the feature representations of the objects. This matrix facilitated the grouping of objects based on their visual similarity. We applied the Constrained K-Means Clustering Algorithm to maintain both cohesive clusters and balanced sizes, ensuring that no single cluster dominated the data distribution. Balanced clusters are crucial for mitigating bias during training and ensuring that the model's performance is not skewed by uneven data splits.

The dataset was divided into 5 folds, with each fold containing 10 objects (as shown in Table 1). This partitioning ensures that each fold is independent while maintaining intra-fold consistency. The use of k-fold cross-validation allows the model to be trained and evaluated on multiple data partitions, promoting a thorough assessment of its generalizability and reducing the risk of overfitting. Within each fold, we identified the 10 inter-class pairs with the highest similarity scores. Additionally, we included intra-class pairs corresponding to each object within the fold, yielding a total of 20 visually similar pairs per fold. This combination of inter- and intra-class pairs ensures that the model encounters a rich variety of learning scenarios.

This structured partitioning strategy is particularly advantageous for class-agnostic visual counting tasks, where the goal is to generalize beyond class labels and respond effectively to diverse visual patterns. By clustering objects based on normalized feature similarity and ensuring balanced fold sizes, we created a robust framework for both training and evaluation. This design not only strengthens the model's ability to learn from diverse data distributions but also ensures reliable performance when generalizing to unseen data in real-world scenarios.

Table 1. FGC200K Dataset Fold Structure

Fold	Categories
Fold 1	bean, bolt, checker, chess, coin, nut, push pin, screw, sunflower seed, washer
Fold 2	berry, button, cashew, catan house, catan money, cracker, pasta, peanut, raisin, scrabble
Fold 3	coffee bean, egg, lemon, m&m, marble, pill, pistachio, poker, pumpkin seed, skittle
Fold 4	binder clip, domino, fork, mahjong, monopoly money, paper star, playing card, puzzle, rice grain, sticky note
Fold 5	beads, bottle, keycap, lego, marker, monopoly house, pen, pencil, plastic cup, screw driver

4. Experiments

4.1. Evaluation Metrics

In alignment with prior work on visual counting, we use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as our primary evaluation metrics. These are defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|c_i - \hat{c}_i|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|c_i - \hat{c}_i|^2}$$

,

where c_i and $\hat{c}_i$ are the ground-truth and predicted count of an object in the i-th image. In this task, each image contains two distinct objects, and the model must estimate the count of both. To handle this, we perform inference on each image twice, once for each object type. This ensures independent predictions for each object, which is crucial given their visual similarity.

Because the two objects are visually similar, the task presents several challenges. A common issue is overcounting, where the model confuses the two objects and sums them together instead of distinguishing them correctly. Alternatively, the model might undercount, failing to detect the objects entirely and predicting a count of zero. These errors are not fully captured by standard metrics such as MAE and RMSE, which focus only on the absolute difference between predicted and true counts. To address this limitation, we propose the Fine-grained Differentiation Metric (FDM), designed specifically to evaluate how well the model differentiates between the two objects.

FDM is defined as:

$$\text{FDM} = \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - \frac{|c_{i,1} - \hat{c}_{i,1}| + |c_{i,2} - \hat{c}_{i,2}|}{c_{i,1} + c_{i,2}})$$

where $c_{i,j}$ and $\hat{c}_{i,j}$ are ground-truth and predicted count of the j-th object in the i-th image.

The FDM metric produces values between 0 and 1. A score close to 0 indicates that the model struggles to differentiate between the two objects, leading to either overcounting or undercounting. Conversely, a score near 1 suggests that the model has successfully distinguished between the two objects and produced accurate counts.

While MAE and RMSE provide a measure of the overall counting error, FDM offers additional insight into the model's ability to correctly differentiate between visually similar objects. This makes FDM particularly valuable in our context, where accurate distinction between the two objects is essential. Together, these metrics provide a more comprehensive evaluation of the model's performance in both object counting and differentiation.

4.2. Quantitative Results

Dataset	MAE ↓	RMSE ↓	FDM ↑
FGC200K	96.05	132.31	0.064
REC-8K (Val set)	11.56	28.02	–
REC-8K (Test set)	11.82	24.98	–
FSC-147 (Val set)	12.14	47.51	–
FSC-147 (Test set)	14.76	120.42	–

Table 2. Comparion of Datasets on the model CountGD (using text only)

As shown in Table 2, our dataset reveals a significant performance gap between class-agnostic and fine-grained counting tasks, with notably higher MAE and RMSE scores. CountGD [1], a state-of-the-art model for open-world counting, performs well on coarse-grained datasets like FSC-147 [11], achieving low MAE and RMSE values. These datasets involve counting distinct object types, where the model can excel using text-only inputs without requiring fine intra-class distinctions. This highlights CountGD's robustness when dealing with broadly defined object categories.

However, the model's performance declines significantly when evaluated on FGC200K, a dataset specifically designed to test fine-grained class-agnostic counting with subtle visual differences. On this dataset, CountGD exhibits much higher MAE and RMSE scores (96.05 and 132.31, respectively), indicating that while it excels in broad categorical counting, it struggles with tasks demanding fine-grained differentiation. The model's reliance on coarse visual features makes it susceptible to over- or undercounting when faced with visually similar objects, limiting its utility in more complex scenarios.

Moreover, the FDM metric further highlights the model's limitations, with a low score of 0.064, indicating frequent instances of both overcounting and undercounting. This result demonstrates that the current model struggles to accurately handle fine-grained scenarios, where subtle distinctions between visually similar objects are crucial. These findings emphasize the need for more advanced counting models capable of overcoming these challenges, ensuring greater precision in complex real-world applications.

4.3. Qualitative Results

In this section, we present the qualitative results derived from analyzing the model's performance across various object counting tasks. The evaluation focused on identifying the model's strengths and weaknesses in distinguishing objects based on visual attributes, including color, shape, and texture.

Overcounting Issues. A thorough examination of the dataset images (Figure 5) reveals a significant overcounting issue, especially when the model encounters visually similar objects. This challenge is expected in fine-grained counting tasks, where subtle visual distinctions complicate differentiation. The model frequently overestimates counts, indicating a need for improved object recognition strategies.

Analysis of intra-class pairs shows a consistent pattern of inaccuracies, with the model either overcounting objects or failing to detect any, resulting in counts of zero. These discrepancies highlight the model's limitations in handling intra-class variations and its struggle to distinguish similar objects within the same category.

Although the model performs slightly better with inter-class pairs, it still tends to overcount, reflecting a lack of understanding of unique features that differentiate object categories. This indiscriminate counting suggests that the model struggles to effectively identify and quantify target classes. Addressing these issues will require further investigation into the model's architecture and training methodologies to enhance its ability to recognize and count objects based on specific features.

Objects Layout and Density. As illustrated in Figure 6, the model's performance is notably impacted by object density within the scene. In cases where objects are closely packed, such as the pasta example, the model struggles to isolate individual items accurately, leading to undercounting. This underestimation occurs despite the need for fine-grained distinctions between different pasta shapes. The dense arrangement causes the model to perceive the collection as a homogeneous mass, which hinders its ability to count accurately. This outcome underscores a critical limitation in the model's capacity to handle complex scenes with high object density, emphasizing the need for improved feature extraction techniques to better capture subtle shape and texture variations.

Text Prompt Variability. The model's reliance on text prompts highlights its limitations in specificity. In our ex-

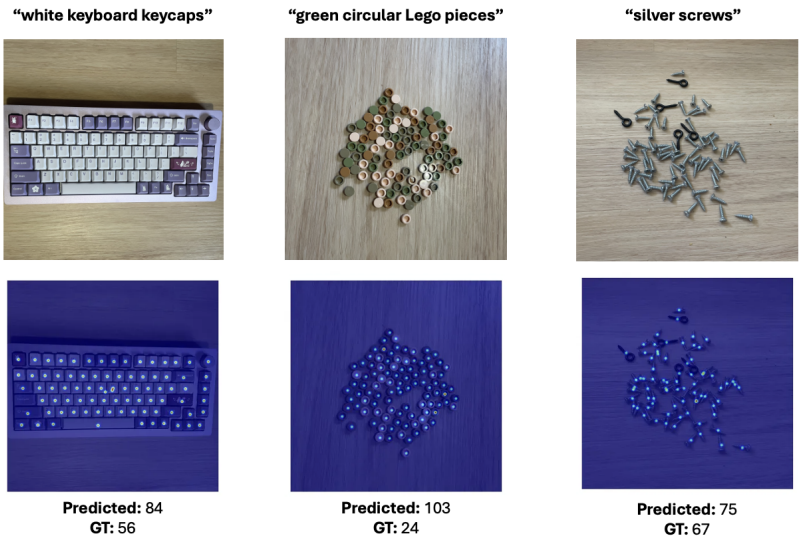

Figure 5. Example tests from the FGC200K dataset demonstrate various attribute distinctions within the same class. Shown are three test pairs: "white vs. purple keyboard keycaps," "green vs. pink circular Lego pieces," and "black vs. silver screws." These pairs illustrate the fine-grained visual differences that the model must distinguish to achieve accurate counting.

Figure 6. Evaluation of model performance on shape and texture distinctions within the same class, exemplified by the pasta pair: spiral vs. penne pasta.

Figure 7. Model performance with a specific prompt: "small, round red poker chips with white markings." For the pair "white poker chips vs. red poker chips"

periments, even highly detailed prompts failed to yield improved counting accuracy. For instance, using a prompt like "small, round red poker chips with white markings" did not prevent the model from indiscriminately counting all chips present in the scene (Figure 7). This indicates a fundamental issue in the model's ability to filter based on visual attributes despite the clarity of the prompts.

Additionally, we tested the model's response to misleading prompts, such as "Large rectangular wooden plates" for a set of circular coins (Figure 8). The model's continued detection and counting of objects unrelated to the prompt reveals a critical flaw in its interpretative capacity. Rather than effectively applying the descriptive constraints, the model appears to rely on generalized recognition patterns. This limitation underscores the necessity for refining prompt alignment mechanisms and the underlying model architecture to enhance specificity in object recognition and counting tasks.

5. Conclusion & Future Work

In this study, we introduced a fine-grained counting task aimed at accurately identifying and counting visually similar objects in real-world scenarios. We developed a novel dataset, FGC200K, comprising 100 distinct categories tailored specifically for this purpose. This dataset not only addresses the intricate challenges of counting in complex environments but also serves as a valuable benchmark for future research. Our proposed evaluation metric provides a com-

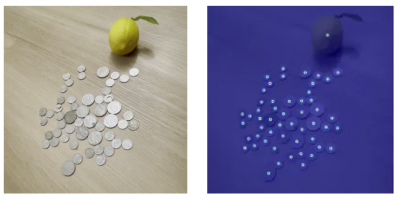

Figure 8. Model response to a misleading prompt: ”Large rectangular wooden plates.” For the pair ”5 cents coin vs. 10 cents coin”

prehensive framework for systematically assessing model performance in fine-grained counting tasks.

Our findings reveal that even state-of-the-art models struggle to accurately count visually similar objects, often over- or under-counting despite detailed prompts. This is due to an overreliance on generalized recognition patterns rather than a nuanced understanding of object attributes. The issue is exacerbated in tests with inter-class variations and misleading prompts, where models struggle to distinguish subtle differences. These results highlight significant gaps in the ability to integrate prompt constraints and interpret context-dependent information, underscoring the need for further research to enhance fine-grained object recognition.

Future Work. Moving forward, we plan to conduct a thorough analysis of the FGC200K dataset to identify potential limitations and areas for improvement. Specifically, we will examine class imbalance, where certain object categories may be underrepresented, and assess variability in object representation to ensure that models trained on the dataset perform consistently across diverse scenarios. We will also explore whether there are subtle overlaps between classes that could confuse models, such as visually similar objects belonging to different categories. Identifying these issues will be crucial for refining the dataset, rebalancing class distributions if necessary, and improving the overall robustness and reliability of the benchmark. Additionally, we intend to expand the dataset further by introducing more challenging cases, such as occluded or partially visible objects, to push model performance beyond its current limits.

On the modelling side, our goal is to develop a specialized counting model using state-of-the-art techniques. Fine-tuning with GroundingDINO [10] will enhance object detection and improve accuracy in complex scenes. To boost generalization, we’ll incorporate synthetic data from recent techniques [9]. We also aim to build a robust counting system that differentiates similar objects using criteria like size, shape, and spatial relationships, while adjusting predictions based on context. Chain-of-Thought (CoT) prompting will enable deeper reasoning, and combining CoT with multimodal learning will enhance interpretive abilities. Finally, we’ll explore incremental learning to ensure the model adapts continuously without retraining, setting new standards for fine-grained counting tasks.

References

[1] A. Amini-Naieni and N. Naieni. Open-world text-specified object counting. *Semantic Scholar*, 2024. 2, 6

[2] Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specified object counting. *arXiv preprint arXiv:2306.01851*, 2023. BMVC 2023. 2

[3] Siyang Dai, Jun Liu, and Ngai-Man Cheung. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024. 3

[4] Y. Dai, Q. Yang, and Z. Li. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4

[6] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network, 2017. 2

[7] Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. *arXiv preprint arXiv:2305.07304*, 2023. 2

[8] Seunggu Kang, WonJun Moon, Euiyeon Kim, and Jae-Pil Heo. VLCounter: Text-aware visual representation for zero-shot object counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024. 2

[9] Lukas Knobel, Tengda Han, and Yuki M. Asano. Learning to count without annotations, 2024. 8

[10] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 8

[11] A. Ranjan, Z. Wu, K. Mo, and J. Malik. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 6

[12] Vishwanath A. Sindagi, Rajeev Yasarla, and Vishal M. Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method, 2020. 2

[13] Jia Wan, Qiangqiang Wu, Wei Lin, and Antoni B. Chan. Robust zero-shot crowd counting and localization with adaptive resolution sam, 2024. 2

[14] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2141–2149, 2021. 2

[15] Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15548–15557, June 2023. 2