IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

When Do Training-Free Counting Methods Fail? A Study on Prompt Sensitivity Using SAM

XINGLIN, ZHONG[1], ARAVIND THIAGARAJAN[2], AYUSH ALPESHBHAI PATEL[3]
[1]Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada (e-mail: azhong@lakeheadu.ca)
[2]Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada (e-mail: athiagar@lakeheadu.ca)
[3]Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada (e-mail: apate245@lakeheadu.ca)

Corresponding author: Xinglin Zhong (e-mail: azhong@lakeheadu.ca).

This project was completed as part of the COMP-5422 Computer Vision and Image Analysis course at Lakehead University.

ABSTRACT In many real-world applications, object counting is used for important tasks, such as taking warehouse inventory, counting products in agriculture, retail, and more. Traditional counting systems often require a lot of manual labour. They take a large amount of time because every object must be manually annotated, which makes the process slow and difficult to scale. The recently published IEEE paper *Training-Free Object Counting with Prompts* proposes a different approach to object counting by using the Segment Anything Model (SAM) and CLIP to perform counting without needing any training or labour-intensive point annotations. In this project, we reproduced the method described in the published paper and evaluated something the original authors did not examine: How reliable the training-free method is when the user provides imperfect prompts or when various types of noise are introduced. In practical situations, a user may draw shifted or inaccurate boxes, place point prompts incorrectly, or enter text descriptions that are misspelled, vague, or completely incorrect. To study this, we injected different types of prompt noise into the model, including spatial shift, scale change, dropped prompts, false prompts, mixed noise, and multiple forms of text noise and applied these noisy prompts to the FSC147 dataset to measure their effect on SAM's accuracy. Our results show that the method remains stable under mild and moderate noise but becomes unreliable when the noise reaches extreme levels, especially with false prompts, incorrect text labels, or strong mixed noise. Overall, our study expands on the findings of the original paper by showing how prompt quality affects the performance of training-free object counting and by identifying where the method holds up and where it fails under real-world noisy conditions.

INDEX TERMS CLIP, computer vision, FSC147 dataset, object counting, prompt noise, SAM, segmentation, training-free methods

I. INTRODUCTION

OBJECT counting in computer vision is an important task in many real-life applications, such as taking inventory in warehouses, tracking products in agriculture and retail, monitoring crowds, and more. Traditional object counting models depend heavily on supervised learning, which requires large amounts of training data and detailed point annotations. This process is very time-consuming and becomes difficult to manage when working with thousands of images [3].

Training-free object counting methods provide a different approach. With a training-free method, the system can determine how many objects are in an image **without** needing labour-intensive data collection and point annotations. This saves time, reduces human error, and makes the method more practical in situations where manual labelling would be slow or impractical.

The recently published IEEE paper *Training-Free Object Counting with Prompts* addresses these challenges by combining the Segment Anything Model (SAM) [4] with CLIP [5] to count objects using simple user-provided prompts such as bounding boxes, point clicks, or text descriptions. This approach removes the need for extensive annotation and avoids the full training pipeline used in traditional models. However, the original paper assumes that user prompts are accurate. In many real situations, prompts may be shifted, incomplete, misplaced, misspelled, or overly general. The original study does not examine how these imperfect prompts affect the reliability of the training-free method.

In this project, we reproduced the method described in the published paper [1] and evaluated its performance under a wide range of prompt noise conditions. Using the FSC147

dataset [2], we introduced spatial shifts, scale changes, dropped prompts, false prompts, mixed noise, and several forms of text noise. By gradually increasing noise levels, we observed where the method remained stable and where it began to fail.

Our results expand on the findings of the original work by showing how sensitive SAM-based counting is to prompt quality. This analysis highlights both the strengths and limitations of the training-free approach and provides insight into how the training-free method behaves under realistic, imperfect input conditions. This leads us to a simple question: *How robust is the SAM-based training-free object counting method when user prompts contain realistic forms of noise?*

II. LITERATURE REVIEW/RELATED WORK

Object counting has been widely studied in computer vision due to its rich applications in various industries, such as agriculture, manufacturing, retail, and crowd monitoring. As mentioned in the introduction, early approaches rely on fully supervised learning, where models are trained to map image regions to density maps (as shown in the vanilla FSC147 dataset) [2] or direct count estimates. Some of the classic supervised methods include multi-column convolutional architectures [3] and regression-based networks. These methods demonstrate high accuracy within their training domains, but they typically struggle to generalize to unseen object categories and require costly annotations.

In order to reduce the labelling man toil, efforts have been made in the advancement towards the few-shot object counting, where only a small number of reference examples are provided per category of the target objects. Feature similarity, or template matching, improvements to adapt to unseen objects with limited supervision. Despite progress, we have found that these "few-shot" models still require a long and potentially costly training stage, and their performance often degrades significantly under certain domain shifts or when used to count objects which are not included, or not similar to the objects that are found in the support set.

More recently, some novel training-free and zero-shot counting techniques emerged, driven by the availability of large vision foundation models. The Segment Anything Model (SAM) is one of the novel methods that exemplifies this area: It enables segmentation of the image using simple prompts such as bounding boxes and points. [4] With the collaboration with other models such as Contrastive Language-Image Pretraining (CLIP), it even accepts text inputs. [5] All without requiring additional object-specific training. This opens a whole new path for object counting, where segmentation for counting can be developed solely through interactive prompts.

However, a key challenge with such prompt-driven methods is their dependency on accurate human input. Unlike fully supervised methods, they learned from a significant amount of labelled data; training-free approaches entrust their accuracy to their users, expecting to receive precise prompts. Unfortunately, in real-world usage, it is common that prompts are frequently imprecise. For instance, bounding boxes may be shifted or resized, points may be misplaced, or even the object may be incorrectly referenced in text. Despite SAM's impressive versatility and efficiency, the robustness of segmentation-based counting under prompt noise remains underexplored.

A small collection of studies has investigated robustness in such models, but primarily under the image level, not annotation errors. While text-guided counting can leverage multiple text-to-image models, their susceptibility to semantic drift, misspellings, or incorrect category descriptions needs to be systematically evaluated further.

Gap Identified: There is a limited amount of literature analyzing how different types and severities of prompt noise affect the accuracy of training-free object counting across diverse object categories. Specifically:

- Spatial inaccuracies (shift/scale errors)
- Incomplete prompts (dropped annotations)
- False guidance (incorrect or irrelevant prompts)
- Textual ambiguity or semantic misalignment

This gap motivates our work, where we systematically evaluate box, point, and text prompt noise across multiple real-world scenarios. The findings provide practical insights into the reliability of prompt-guided foundation models and highlight failure modes that are overlooked in prior counting methods.

III. METHODOLOGY

A. DATASET

We evaluate our training-free prompt-guided counting approach on the FSC147 dataset [2], a widely used benchmark for few-shot and zero-shot object counting. Our FSC147 contains 6146 images covering 147 diverse object categories, including fruits, tools, household items, packaged goods, and natural objects. Each image is annotated with point-level ground truth representing the center of every object instance.

In contrast to conventional studies that use all training and validation splits, our experiments focus on random subsets of 100 images drawn from the test split. The reasons for that are, but not limited to:

- Prior work done by the previous researchers emphasizes the performance of their project, but we are focusing on how robust it is.
- Limitations on our hardware preclude an extremely large dataset test.
- Our inferential expectation is that the results for images containing small and same objects should outperform those of large and different objects.

Therefore, we have set up our randomized image selectors to pick:

- 30 random images that contain less than or equal to 20 objects,
- 40 random images that contain between 21 and 50 objects, inclusive,
- 30 random images that contain over 50 objects.

Consideration of the CARPK Dataset

We initially considered using the CARPK dataset [6]; however, we ultimately decided not to include it for two reasons. First, CARPK contains only car images, which limits the alignment with our goal of evaluating prompt robustness across diverse object categories. Second, our prompt-noise experiments require repeated SAM and CLIP inference, and running these tests on multiple datasets would exceed our available computational resources. For these reasons, we restricted our experiments to FSC147. However, it is open for future experiments with higher computational resources available.

B. MODEL AND COUNTING PIPELINE

We replicated the outcome using the original paper's method [1]. It is built entirely on top of the SAM and CLIP [4] [5], a segmentation model capable of generating object masks from simple user-provided prompts. What is more important, SAM is not trained or fine-tuned on FSC147, and no task-specific parameters are learned. In fact, SAM has never encountered any data in the FSC147.

Prompting Modes

We investigate three prompts that are available for this project:

- Bounding box prompts: a rectangular region specifying the targeted object.
- Point prompts: the center of the targeted object derived from the box annotations.
- Text prompts: a natural language object description passed through CLIP and to SAM.

Segmentation and Counting

Given an input prompt, SAM produces a set of candidate masks. We filter these masks by confidence and instance separation, then compute the number of detected (predicted) objects as:

$$\hat{C} = \sum_{i=1}^{N} \mathbf{1}\{\text{mask}_i \text{ is valid}\}. \tag{1}$$

This ''segment-and-count'' strategy is fully training-free. No density maps or regression heads are used, even though they are included in the FSC147 dataset for reference.

C. PROMPT NOISE MODELING

Real-world users rarely provide perfect prompts. To examine the practical reliability of this training-free counting method, we simulate controlled noise across all three prompt modalities: box, point, and text. For each experiment, the noise type and parameters are applied independently per image.

1) Box Prompt Noise

Bounding boxes are modified in four different ways:

a: Shift Noise

The bounding box is translated randomly:

$$(x_1', y_1') = (x_1 + \Delta x,\ y_1 + \Delta y), \qquad (x_2', y_2') = (x_2 + \Delta x,\ y_2 + \Delta y). \tag{2}$$

We examine several shift magnitudes: ± 4, ± 8, ± 15 pixels.

b: Scale Noise

Box width and height are scaled by multiplication:

$$w' = sw, \qquad h' = sh. \tag{3}$$

We test both enlargement and shrinkage at $\pm 10\%$ and $\pm 30\%$.

c: Drop Noise

A random subset of bounding boxes is removed entirely. We evaluate drop probabilities of 0.1, 0.3, 0.5.

d: False Box Noise

Extra bounding boxes are added at random background locations. We test 10%, 20%, and 40% false boxes relative to the true count.

e: Mixed Box Noise

A combination of shift + drop + false boxes using "mild," "medium," and "maximum" settings as defined in the `main-fsc147` file. This simulates real user mistakes where multiple annotation errors occur simultaneously.

2) Point Prompt Noise

Point-based annotations typically represent object centers. We introduce analogous noise types to the box prompt type, with the absence of scale noise.

a: Shift Noise

Point coordinates are shifted by ± 3, ± 6, ± 8 pixels.

b: Drop Noise

Correct points are randomly omitted at probabilities 0.1, 0.3, 0.5.

c: False Point Noise

Additional false points are inserted in the background at levels 10%, 20%, 40% of the true count.

d: Mixed Point Noise

We define "mild," "medium," and "max" mixtures (e.g., ± 3 px + 10% drop + 10% false, etc.) to model realistic annotation inconsistencies.

3) Text Prompt Noise

Text prompts represent semantic descriptions of the target object and are susceptible to language noise. We divided this into four categories:

- Mild Noise: adjectives or modifiers added without changing object identity.
- Misspellings: typos inserted into the prompt.
- Related-Class Prompts: superordinate or category-level descriptions.
- Wrong Prompts: totally incorrect object to count, to test model hallucination.

We also analyze the extreme case where the model receives an incorrect prompt applied to images where the target count is nearly zero.

D. EVALUATION METRICS

To measure counting performance under noise, we adopt four standard metrics, as the original research paper demonstrated, but with some modifications:

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- NAE (Normalized Absolute Error)
- SRE (Squared Relative Error)

These metrics align with prior FSC147 literature and allow fair comparison across object categories with varying densities and sizes. All results are computed per image and averaged over the test set.

For experiments involving Wrong Prompts in text prompt noise, the ground truth counts will be zero because the prompt does not correspond to any object in the image. Metrics such as NAE and SRE include division by the ground-truth count, which becomes undefined when the denominator is zero. To avoid division by zero errors and ensure numerical stability, we replaced the ground-truth count with a small constant:

$$C_i = 0.001. \tag{4}$$

Now, we are expecting very large NAE and SRE when the model miscounts in the Wrong Prompts noise scenario.

E. IMPLEMENTATION DETAILS

Model: Meta AI's SAM model, OpenAI's CLIP model for text to image.

Hardware: GPU with CUDA support

Post-processing: Small connected components removed; mask overlaps resolved using non-maximum-area suppression; false prompt regions filtered via thresholding.

Reproducibility: All noise transformations use fixed random seeds; parameters reported exactly match the settings used in the test results file.

IV. RESULTS

This section presents the performance of the training-free counting system under various types of prompt noise. For each experiment, we report MAE, RMSE, NAE, and SRE. Each table or figure is immediately followed by its descriptive text summarizing the observed behaviour.

TABLE 1: Box shift noise results.

Noise	MAE	RMSE	NAE	SRE
Clean	12.59	22.12	0.36	3.07
Shift 4	12.99	22.95	0.37	3.12
Shift 8	14.26	33.90	0.36	3.32
Shift 15	16.50	38.01	0.43	3.96

The results in Table 1 show that the model tolerates small spatial misalignment ($\pm$4 px) with only minor increases in error. Larger shifts ($\pm$15 px) cause noticeable degradation, indicating that SAM's segmentation becomes unreliable when the prompt no longer overlaps the correct region.

TABLE 2: Box scale noise results.

Noise	MAE	RMSE	NAE	SRE
Clean	12.59	22.12	0.36	3.07
Scale 0.1L	12.66	22.22	0.37	3.11
Scale 0.1S	12.75	22.17	0.38	3.22
Scale 0.3L	12.81	22.33	0.38	3.17
Scale 0.3S	12.63	22.82	0.38	3.54

Table 2 (L means enlarge, S means shrink) demonstrates that modest scaling ($\pm$10%) has minimal effect on accuracy. Extreme scaling ($\pm$30%) increases error because enlarged boxes include unrelated background, while shrunk boxes fail to fully cover objects.

TABLE 3: Box drop noise results.

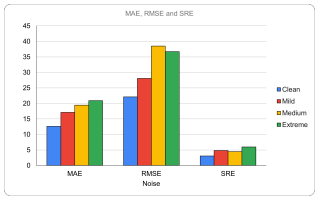
Noise	MAE	RMSE	NAE	SRE
Clean	12.59	22.12	0.36	3.07
Drop 0.1	12.30	21.23	0.36	3.06
Drop 0.3	12.04	20.69	0.35	2.95
Drop 0.5	15.54	32.42	0.39	3.47

As seen in Table 3, dropping a small proportion of box prompts sometimes slightly improves RMSE by reducing clutter. However, removing half of the boxes leads to a strong performance drop.

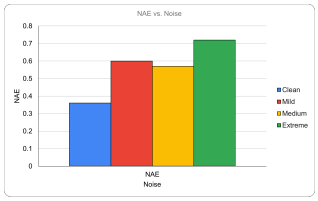
TABLE 4: Box false-noise results.

Noise	MAE	RMSE	NAE	SRE
Clean	12.59	22.12	0.36	3.07
False 0.1	16.34	27.32	0.58	4.59
False 0.2	16.74	28.33	0.58	4.88
False 0.4	17.19	28.47	0.59	4.92

Table 4 shows that false box prompts consistently cause overcounting. SAM tends to interpret all user-provided boxes as meaningful, making this noise type particularly damaging.



(a) MAE, RMSE, SRE vs. mixed box noise.



(b) NAE vs. mixed box noise.

FIGURE 1: Performance under mixed box prompt noise.

Figure 1 illustrates that mixed noise produces the strongest degradation among all box prompt conditions. Mild mixed noise is manageable, but medium and extreme levels cause sharp increases across all four error metrics.

TABLE 5: Point shift noise results.

Noise	MAE	RMSE	NAE	SRE
Clean	12.37	22.33	0.34	2.85
Shift 3	12.48	22.34	0.35	2.96
Shift 6	12.20	21.36	0.35	2.88
Shift 8	12.88	23.20	0.36	3.11

Table 5 shows that point prompting is robust to moderate spatial shifts. Only the ± 8 px condition produces a noticeable increase in error.

TABLE 6: Point drop noise results.

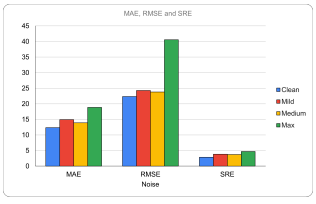
Noise	MAE	RMSE	NAE	SRE
Clean	12.37	22.33	0.34	2.85
Drop 0.1	12.15	22.09	0.33	2.78
Drop 0.3	12.24	22.96	0.33	2.78
Drop 0.5	13.82	29.83	0.33	2.98

As shown in Table 6, dropping a small portion of point prompts has limited effect, whereas dropping half of them leads to clear increases in MAE and RMSE.

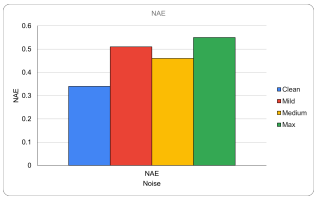
TABLE 7: Point false-noise results.

Noise	MAE	RMSE	NAE	SRE
Clean	12.37	22.33	0.34	2.85
False 0.1	15.38	25.55	0.50	3.88
False 0.2	13.90	23.18	0.48	3.70
False 0.4	26.20	46.36	0.84	6.66

Table 7 demonstrates that false point prompts are highly disruptive. Even moderate false-point ratios cause substantial overcounting.



(a) MAE, RMSE, SRE vs. mixed point noise.



(b) NAE vs. mixed point noise.

FIGURE 2: Performance under mixed point prompt noise.

Figure 2 shows that the mixed-noise condition amplifies errors much more strongly than shift, drop, or false noise individually. The max mixed setting yields the worst overall results for point prompting.

TABLE 8: Mild text noise results.

Condition	MAE	RMSE	NAE	SRE
Clean	16.63	31.73	0.37	3.71
Mild Noise	16.39	29.70	0.38	3.72

Table 8 shows that simple descriptive variations (e.g., adjective modifiers) do not significantly affect CLIP-based performance.

TABLE 9: Related / generalized text prompt results.

Condition	MAE	RMSE	NAE	SRE
Clean	16.63	31.73	0.37	3.71
Related / Generalized	19.14	34.56	0.44	3.97

Table 9 shows that substituting the target word with a more general category (e.g., "fruit") increases error.

TABLE 10: Misspelled text prompt results.

Condition	MAE	RMSE	NAE	SRE
Clean	16.63	31.73	0.37	3.71
Misspelled	19.89	36.66	0.46	4.39

Table 10 indicates that even minor misspellings degrade CLIP embeddings enough to harm segmentation performance.

TABLE 11: Wrong-class text prompt results.

Condition	MAE	RMSE	NAE	SRE
Clean	16.63	31.73	0.37	3.71
Wrong Class	29.50	59.63	0.54	4.89

Table 11 shows the severe degradation caused by incorrect semantic prompts. CLIP assigns embeddings for the wrong class, leading SAM to segment irrelevant or nonexistent regions.

TABLE 12: Wrong-class results with $gt\_cnt = 0.001$.

Condition	MAE	RMSE	NAE	SRE
Clean	16.63	31.73	0.37	3.71
gt_cnt=0.001	27.02	42.84	27019.08	1354.67

Finally, Table 12 highlights the instability of NAE and SRE under wrong prompts when ground-truth counts approach zero. The semantic misalignment results in extremely large relative errors.

V. DISCUSSION

The experimental results reveal clear patterns of the robustness of this novel training-free object counting method under different forms of prompt noise. Overall, the method demonstrates moderate tolerance to small spatial inaccuracies but is highly sensitive to misleading or incorrect semantic guidance, particularly when using text prompts.

A. DISCUSSION ON BOX PROMPT NOISES

The model handles small spatial shifts and scale variations with limited degradation, indicating that SAM does not require perfectly precise bounding boxes if the prompt remains reasonably aligned with the target region. However, larger variations in such a field significantly increase error, suggesting that once the box moves too far from the objects, the segmentation becomes unstable.

Drop noise shows a mixed effect: removing a small number of boxes occasionally reduces error by decreasing clutter, whereas removing many boxes consistently harms performance because the model loses necessary spatial cues.

False box prompts cause the most severe performance degradation in this category, as SAM tends to interpret any user-provided box as a meaningful object region, resulting in consistent overcounting, which is a weakness we have found in our experiment.

B. DISCUSSION ON POINT PROMPT NOISES

Point prompting exhibits similar trends but generally shows stronger robustness. Small point shifts have minimal impact, revealing that SAM is being tolerant of slight inaccuracies in point placement. However, performance deteriorates when too many points are removed or when false points are introduced. The mixed point noise experiments highlight that combining several small errors leads to compounded degradation, though still less severely than with box prompts.

C. DISCUSSION ON TEXT PROMPT NOISES

Text prompt noise produces the most desperate decline in this method's performance. Since text prompting in this project relies very heavily on CLIP, the system is heavily dependent on CLIP's ability to correctly interpret the semantic meaning of the text prompt. Mild variations such as descriptive adjectives do not significantly affect performance, since it does not change the fundamental description and meaning of the targeted object.

However, even small misspellings, related class terms, or synonyms start to distort the CLIP performance in its embedding process. Consequently, this leads to segmentation failures when SAM receives an embedding that does not accurately describe the visual target. Wrong-category prompts result in the largest errors across the entire experiment, as CLIP "confidently" produces an embedding for a completely incorrect object class, and since SAM has no prior knowledge of the targeted object, it also "confidently" segments irrelevant or nonexistent regions.

After applying a small constant (gt_cnt = 0.001) to stabilize the ground-truth count, NAE and SRE become extremely large. This exaggerates the illustration of how disruptive semantic misalignment is for CLIP-based text prompting.

D. COMPARISON ACROSS PROMPT TYPES

Comparing the three prompt noises, box and point prompts are significantly more robust than text prompts, where the point prompt is generally the most stable, with boxes moderately robust but prone to errors when false prompts are introduced. Text prompts degrade rapidly even under mild semantic noise, revealing a fundamental vulnerability of CLIP-driven text-to-image guidance: its embeddings vary unpredictably when the input text is incomplete, ambiguous, or incorrect.

E. OVERALL OBSERVATION

Overall, these findings highlight that training-free, prompt-guided counting highly entrusts its performance to the accuracy and clarity of the user's prompt. Box and point prompts can be tolerated to some degree, but semantic errors, especially those involving CLIP text embeddings, will lead to rapid and severe performance degradation. This suggests that real-world use of such a system would benefit from good prompt validation and better text embedding filtering.

VI. CONTRIBUTIONS

Our work provides a systematic investigation into the robustness of training-free, prompt-guided object counting, focusing on how different types of prompt errors affect segmentation-based counting performance. Unlike prior studies that emphasize accuracy under ideal prompting conditions, our work highlights practical failure modes that may occur when prompts are imperfect, ambiguous, or semantically incorrect. The key contributions of this study are as follows:

- **A comprehensive evaluation of prompt noise across three prompts.** We analyze bounding box prompts, point prompts, and CLIP-based text prompts under multiple noise categories, including spatial shifts, scale distortions, dropped prompts, false prompts, and combined mixed noise. This provides a unified and systematic comparison rarely addressed in previous work.
- **Robustness study of CLIP-driven text prompts for object counting.** While text-based prompting is increasingly used, its vulnerability to misspellings, semantic broadening, and incorrect class descriptions has not been quantitatively examined by previous works. Our results expose CLIP's significant fragility when text input deviates from the target object.
- **Identified the distinct failure modes for training-free counting.** We show that spatial inaccuracies (box/point shifts) are generally tolerated, whereas false prompts and semantic errors lead to severe degradation in performance. These findings clarify which types of user errors are most detrimental in real-world applications.
- **Mixed-noise evaluations that reflect realistic user behaviour.** By combining several noise types (shift + drop + false), we expose compounding effects that push the counting system into unstable regimes. This mixed-noise evaluation represents a more realistic scenario than isolated noise cases.
- **Reproducible evaluation pipeline.** All noise processes, random seeds, and evaluation metrics follow a structured pipeline that can be reused for future studies on prompt-based computer vision models or applied to other segmentation-based counting methods.

VII. CONCLUSION

To conclude, our project reproduced the training-free object counting method proposed in *Training-Free Object Counting with Prompts* and evaluated how well it performs when user prompts are noisy or imperfect. Unlike the original study,

which only used accurate and clean prompts, our work examined several types of real-world prompt noise, including spatial errors, missing annotations, false prompts, and different forms of text-based noise. All experiments were conducted on the FSC147 dataset, and each noise type was tested at multiple intensity levels to observe how long the method remained reliable and when it began to fail.

Our results show that SAM-based counting is generally robust when exposed to small spatial errors in box and point prompts. Moderate levels of shift, scale change, or slight point inaccuracies only caused mild changes in accuracy. However, extreme levels of noise, such as large spatial distortions, high amounts of dropped prompts, and especially false prompts, consistently reduced performance. The method reached failure most clearly under mixed noise, where several small mistakes occurred together and created strong degradation across all metrics. Among all prompt types, text prompts were the most sensitive. Even small misspellings or general category names noticeably reduced accuracy, and wrong-class prompts produced severe errors.

Through this study, we identified the conditions where training-free counting remains stable and where it breaks down. These findings highlight the current limitations of SAM and CLIP when used for training-free object counting. We hope that our experiments provide useful insight for future improvements, as understanding the model's weaknesses is an important step toward making training-free counting more reliable in real-world applications.

ACKNOWLEDGMENT

We are greatly thankful to our supervisor, Dr.Garima Bajwa, for her unbounded support and guidance. We are grateful to Zenglin Shi, Ying Sun, and Mengmi Zhang, original authors of the research paper Training-free Object Counting with Prompts, for their hard work.

REFERENCES

[1] Z. Shi, Y. Sun, and M. Zhang, "Training-free Object Counting with Prompts." Available: https://openaccess.thecvf.com/content/WACV2024/papers/Shi_Training-Free_Object_Counting_With_Prompts_WACV_2024_paper.pdf

[2] X. C. Ngo, "FSC147 Dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/xuncngng/fsc147-0 (accessed Dec. 11, 2025).

[3] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in *Proc. CVPR*, 2016. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/papers/Zhang_Single-Image_Crowd_Counting_CVPR_2016_paper.pdf

[4] "Introducing Meta Segment Anything Model 3 and Segment Anything Playground," Meta.com, 2022. https://ai.meta.com/blog/segment-anything-model-3/ (accessed Dec. 11, 2025).

[5] OpenAI, "CLIP: Connecting text and images," Openai.com, 2021. https://openai.com/index/clip/

[6] "CARPK Dataset-Machine Learning Datasets, "Machine Learning Datasets, Jun. 13, 2023. [Online]. Available: https://datasets.activeloop.ai/docs/ml/datasets/carpk-dataset/ (accessed Dec. 11, 2025).