Visual Association Based Zero-Shot Object Counting

Author: Nitin Yadav

Abstract

Zero-Shot Object Counting (ZSOC) aims to estimate the number of object instances specified solely by a textual query, without requiring class-specific training examples. A major challenge in ZSOC lies in the quality of visual exemplars used to establish visual–textual associations—noisy, multi-object, or background-heavy patches can significantly degrade performance. This project replicates the Visual Association-based Zero-Shot Object Counting (VA-Count) framework **[2]**, which tackles this issue through a two-stage pipeline. The first stage, the Exemplar Enhancement Module (EEM), employs GroundingDINO **[3]** and a CLIP-based classifier to automatically extract clean, single-object exemplars while filtering out noisy proposals. The second stage, the Noise Suppression Module (NSM), uses these refined exemplars to supervise a contrastive density-regression model for robust, class-agnostic counting. In this work, I reconstruct the full data-generation and bootstrapping pipeline for exemplar extraction, implement the binary classification module, and generate the training datasets required for NSM. Quantitative results align with expected trends in zero-shot counting, and qualitative visualizations confirm that exemplar-guided contrastive learning remains effective. To support reproducibility, I aim to release the code, preprocessed exemplars, and trained model weights.

1. Introduction

Zero-Shot Object Counting (ZSOC) aims to estimate the number of object instances belonging to an arbitrary category specified only by a textual query at inference time. Unlike traditional counting approaches that rely on category-specific supervision, ZSOC requires models to generalize to entirely unseen object classes, making it a challenging and increasingly important problem in computer vision. State-of-the-art ZSOC methods typically rely on *visual exemplars*—small image patches representing the target object—to establish a visual–textual association. However, the quality of these exemplars is critical: patches with occlusions, cluttered backgrounds, or multiple object instances can severely degrade the final counting performance.

This project focuses on replicating and validating the Visual Association-based Zero-Shot Object Counting (VA-Count) framework **[2]**, which directly addresses the exemplar-quality bottleneck. The framework introduces a two-stage architecture designed to generate and exploit high-quality exemplars in a fully automated manner.

(1) Exemplar Enhancement Module (EEM)
The EEM automatically discovers potential exemplar patches using the open-vocabulary detector GroundingDINO **[3]** and then filters them with a CLIP-based single-object classifier **[11]**. This removes noisy, ambiguous, or multi-object patches, ensuring that only clean exemplars guide the counting model.

(2) Noise Suppression Module (NSM).

The NSM uses the refined positive and negative exemplars to supervise a contrastive density-regression architecture trained on FSC-147 **[1]**. By enforcing alignment between exemplar-guided density maps and ground-truth density fields, the NSM learns a robust, class-agnostic counting mechanism capable of handling distractors, clutter, and background confusion.

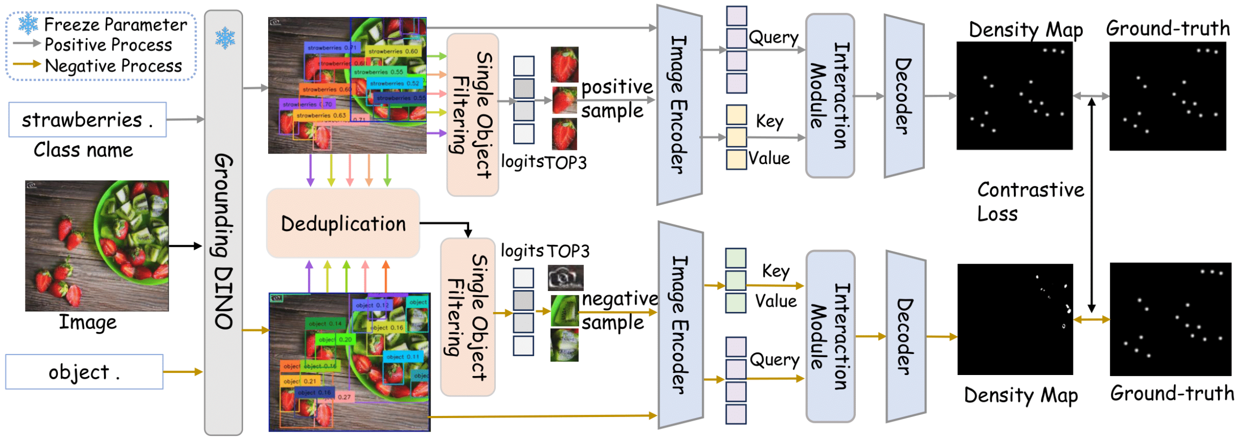


Figure 1: VA-Count framework. GroundingDINO provides candidate proposals, which are deduplicated and filtered by a CLIP-based single-object classifier to yield positive and negative exemplars (EEM). These exemplars guide density-map prediction through an interaction module and dual-branch decoder (NSM), supervised by density regression and contrastive losses.

2. Background and Related Work

Object counting has progressed from class-specific density-map regression toward modern class-agnostic and zero-shot paradigms. Early approaches such as CSRNet **[4]** and related density-map models learned to predict continuous object density fields whose spatial integral yields the total count. These methods achieved strong performance on crowd datasets but required category-specific supervision and lacked generalization to unseen objects. To address scalability, few-shot and class-agnostic counting frameworks emerged. Learning to Count Everything (LCE) **[1]** introduced the exemplar-matching paradigm, where the model uses a small set of annotated exemplar patches to guide counting in novel categories. LCE introduced the FSC-147 dataset with dot and box annotations for few-shot counting.The VA-Count framework leverages these dot annotations to construct ground-truth density maps, which are essential for training its contrastive exemplar-guided density regression model to generate exemplar patches and remove dependency on manual annotation **[2]**.

Recent progress in zero-shot object counting (ZSOC) leverages advances in vision–language modeling and open-vocabulary detection. Models such as GroundingDINO **[3]** demonstrate strong text-conditioned localization by aligning language embeddings with region proposals, while CLIP **[11]** provides robust image–text alignment for semantic filtering. Recent ZSOC methods combined these tools to replace manually curated exemplars with text-driven supervision, as seen in CLIP-Count **[8]**. Most recently, multimodal foundation models such as MolMo **[9]** demonstrate that high-quality point-based datasets (e.g., PixMo-Points **[10]**; see appendix for details) can further generalize class-agnostic counting. Point-level supervision—collected efficiently through human–AI collaboration—provides precise grounding signals that eliminate the need for handcrafted exemplar mining and a large amount of highly precise data can be collected with small human efforts. These models show that object counting and pointing can emerge naturally when trained on dense multimodal reasoning traces, marking a shift toward data-centric open-vocabulary counting.

VA-Count **[2]** addresses this challenge with a two-stage approach: an Exemplar Enhancement Module (EEM) that automatically discovers and filters single-object patches via GroundingDINO proposals and a CLIP-based binary classifier, followed by a Noise Suppression Module (NSM) that uses positive and negative exemplars in a contrastive density-map framework. This design significantly improves robustness to noisy exemplars and distractor objects.

This project contributes to this evolving landscape by not only replicating VA-Count's exemplar-enhancement and density-based counting modules **[2]**, but also by aiming to release clean, structured datasets, code, and trained model weights. These artifacts improve reproducibility and provide an open foundation for future research in zero-shot object counting and multimodal grounding.

3. Data and Method

3.1 Datasets

The FSC-147 dataset is used for training the zero-shot object counting model. This dataset contains 8,502 full-resolution images (384×384) spanning 147 object categories and includes dot annotations marking every object instance. It also provides train/validation/test splits constructed over **non-overlapping classes**, enabling rigorous zero-shot evaluation. For supervision, the ground-truth density maps from the original LCE paper are utilized, wherein the dot annotations are transformed into continuous density fields using Gaussian kernels. These density maps serve as the target signal for the contrastive exemplar-guided density regression.

3.2 Pretrained Models and Reused Code

To maintain computational feasibility while preserving the two-stage structure of VA-Count, the implementation incorporates several pretrained components. **GroundingDINO-T** is used to generate both class-specific and universal object proposals; this model remains frozen throughout and provides consistent region proposals for exemplar selection. For exemplar classification, a frozen CLIP ViT-B/16 encoder is employed as a robust visual feature extractor, enabling effective discrimination between single-object and multi-object patches. The Noise Suppression Module (NSM) is constructed upon a CSRNet/CounTR-based backbone pretrained for dense prediction tasks; this encoder is fine-tuned during training to allow exemplar-guided density estimation to be learned. These pretrained components provide strong initialization and stability, permitting the key contributions of the VA-Count pipeline.

3.3. Method

3.3.1 Exemplar Enhancement Module (EEM)

The Exemplar Enhancement Module (EEM) aims to automatically extract reliable **positive** and **negative** exemplars from each image, addressing the core challenge that FSC-147 does not provide patch-level labels. Following VA-Count, the EEM workflow was streamlined using components from the original VA-Count codebase to reconstruct the pipeline from scratch.

GroundingDINO generates two sets of proposals:

- **Class-specific proposals** using the target object prompt (e.g., *"strawberries"*), and
- **Universal proposals** using a broad prompt (*"object"*), which serve as potential negatives.

All proposed boxes are cropped and stored for further processing. Because universal prompts frequently return boxes overlapping with true object regions, **IoU-based deduplication** (threshold = 0.5) is performed for removing any negative patch that significantly overlaps with a positive one. This ensures a clean negative pool free of accidental true positives. To filter out multi-object or low-quality patches, a binary single-object classifier is trained on features extracted by a frozen CLIP ViT-B/16 encoder and processed through a lightweight two-layer FFN (Figure 3). The training data is produced from two automatically generated sources:

- **/annotated_images**: verified or algorithmically clean single-object patches
- **/ annotated_images_n**: random multi-object crops and noisy proposals

Finally, for each image, **Top-3 positive** and **Top-3 negative** exemplars are selected based on a weighted combination of GroundingDINO logits and classifier scores. These curated exemplars function as robust visual queries for the Noise Suppression Module.

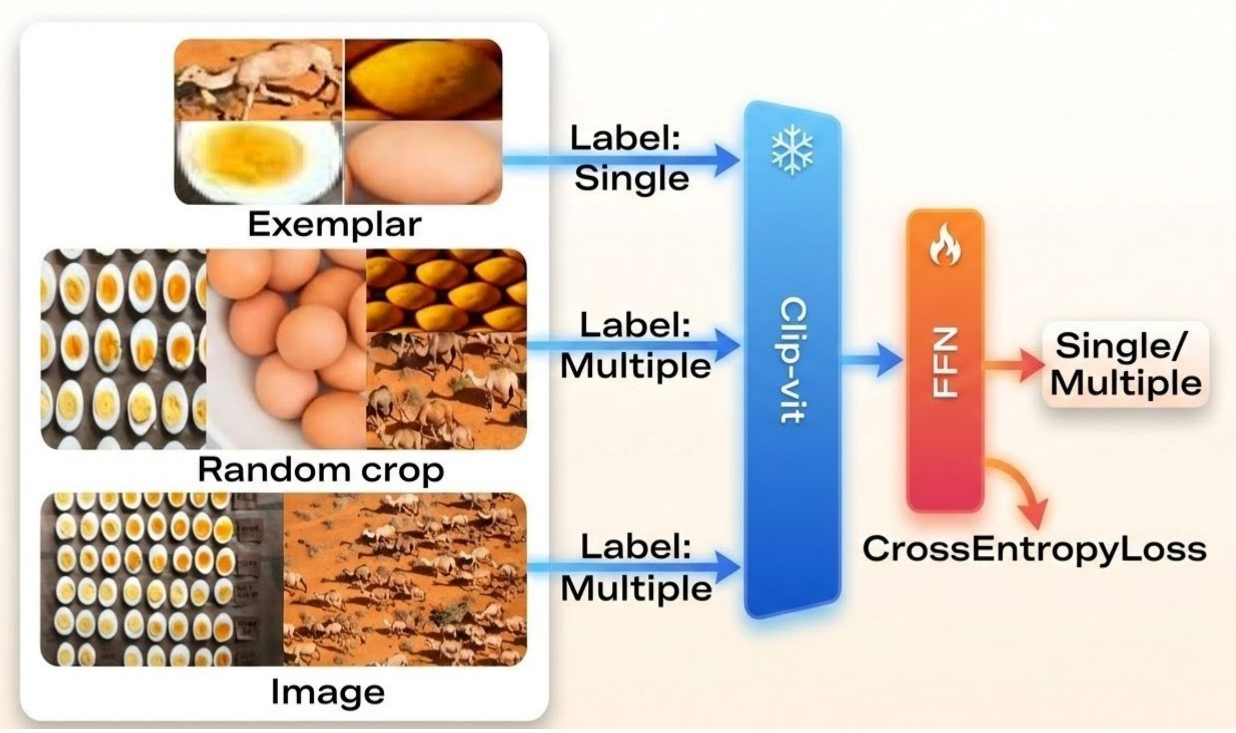


Fig. 3: Illustration of the single object exemplar filtering with a frozen Clip-vit encoder and a trainable FFN to distinguish single from multiple objects.

3.3.2 Noise Suppression Module (NSM)

The Noise Suppression Module (NSM) uses the EEM-selected exemplars to guide density-map estimation through contrastive learning. This implementation mirrors VA-Count's design while incorporating FSC-147's precomputed ground-truth density maps. The input image and its Top-3 positive and negative exemplars are encoded with a shared **CounTR-based convolutional backbone**. Image features act as **Query** tokens, while exemplar features form **Key–Value** tokens. These are fused using a lightweight **interaction module** inspired by CounTR's attention mechanism:

- **Positive exemplars** amplify target-object regions.
- **Negative exemplars** emphasize distractors and background structures.

The fused representation is passed through a CSRNet/CounTR-style decoder to generate two outputs:

- D_p: density map guided by positive exemplars
- D_n: density map guided by negative exemplars

Training follows a two-part loss:

- **Density regression (MSE):** aligns $D_p$with the ground truth

- **Contrastive loss:** pushes $D_p$closer to $D_g$while encouraging $D_n$to diverge

This dual objective enables the model to suppress distractors and visually similar non-target regions, yielding cleaner and more reliable density maps.

At inference time, only the **positive branch** is used, and the final count is simply the integral of D_p:

$$\text{Count} = \sum_{x,y} D_p(x, y)$$

4. Experiments and Results

This section evaluates the replication of the VA-Count framework using both quantitative metrics and qualitative analyses. FSC-147 is used as the primary benchmark and examine model behavior through loss curves, MAE/RMSE trajectories, exemplar quality, and density-map predictions. It is important to note that, due to limited computational resources, the models were pretrained for only 100 epochs and fine-tuned for 125 epochs, whereas the original VA-Count authors report 500 epochs of pretraining and 1000 epochs of full training. Despite this significantly reduced training budget, the results still reflect the characteristic learning dynamics and qualitative behaviors expected from exemplar-guided class agnostic counting.

4.1 Training Dynamics

Optimization behavior was monitored throughout training using Weights & Biases. The trend reflects oscillating loss but overall downward indicating convergence in fewer epochs.

- **Training Loss (Figure 4.1 right):**
  The loss curve shows large oscillations in the early stages, especially before ~15k steps, reflecting high exemplar variability. Once the exemplar filtering becomes more reliable, the loss sharply reduces and stabilizes, with significantly lower variance in the latter half of training.
- **Learning Rate (Figure 4.1 left):**
  The LR schedule follows the expected warm-up to peak phase, followed by a cosine decay across the remaining steps.

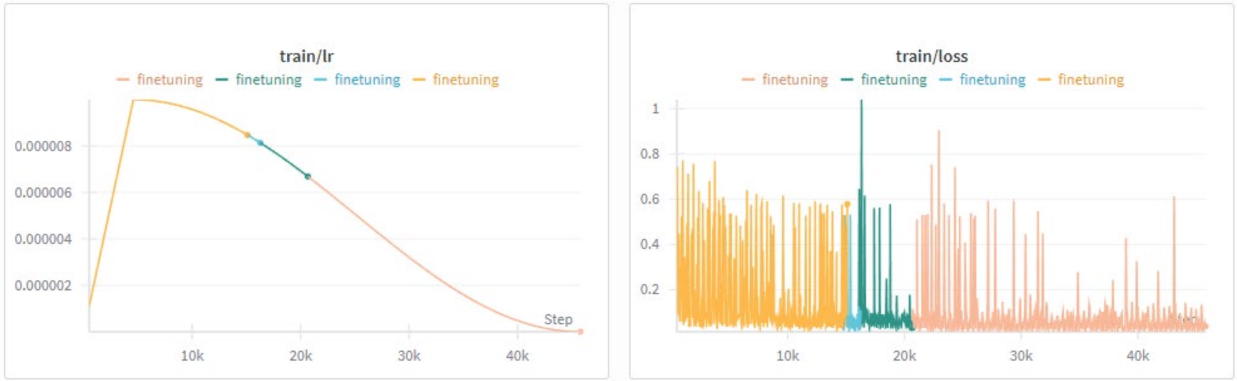


Figure 4.1: Learning-rate (left) and training-loss (right) curves. The LR follows a warm-up and cosine-decay pattern, while the loss becomes progressively smoother as exemplar filtering stabilizes.

4.2 Qualitative Results

4.2.1 Exemplar Quality (EEM Output)

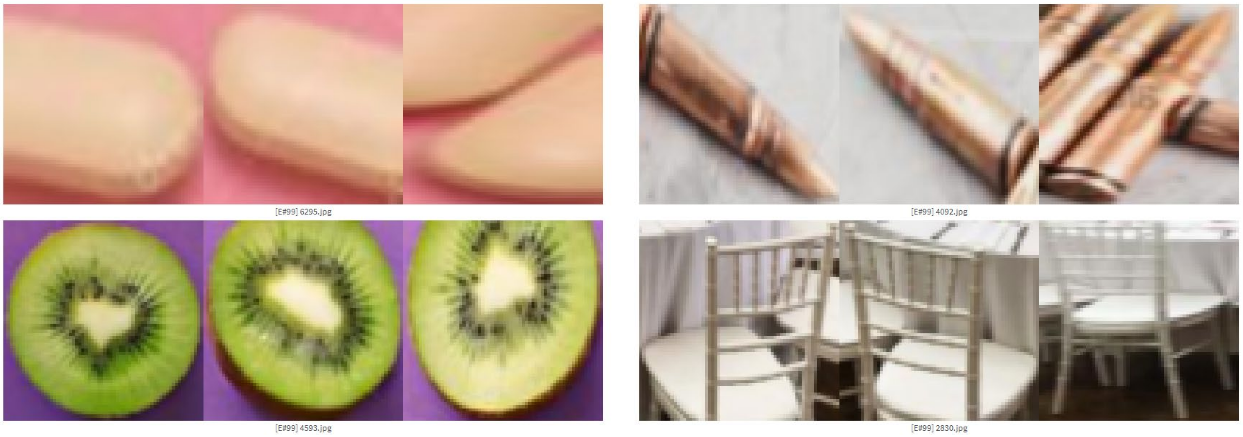


Figure 4.2.1: Illustration of the final positive exemplars for images on FSC-147.

4.2.2 Density-Map Predictions

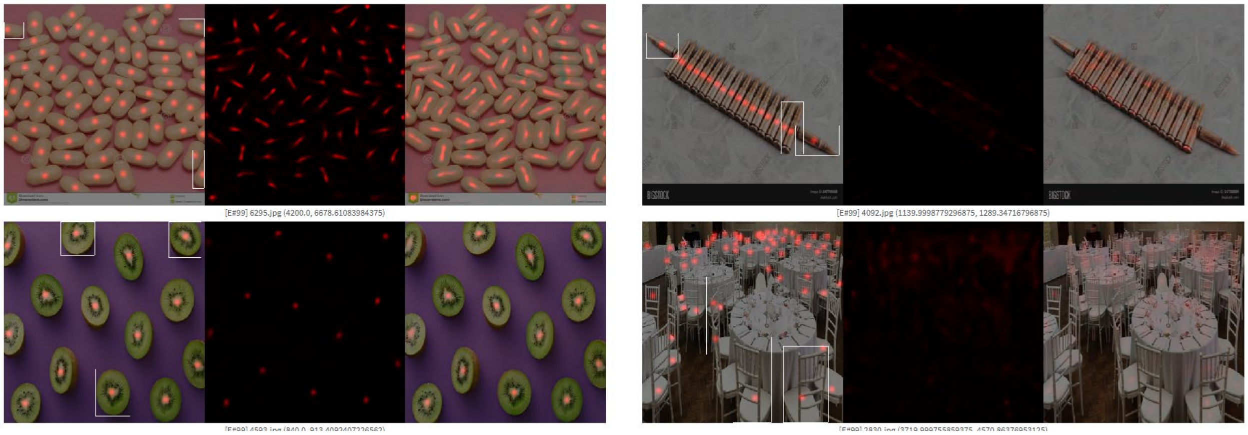

[E#99] 6295.jpg (4200.0, 6678.61083984375)
[E#99] 4092.jpg (1139.9998779296875, 1289.34716796875)
[E#99] 4593.jpg (840.0, 913.4092407226562)
[E#99] 2830.jpg (3719.999755859375, 4570.86376953125)

Figure 4.2.2: Predicted vs. Ground-Truth Density Maps. White boxes show the top 3 exemplars of single object.

4.3 Quantitative Results

The evolution of MAE and RMSE during training and validation is reported in Table 1 and illustrated in figure 4.3 below. Both train MAE/RMSE exhibit substantial early fluctuations, driven by noisy or multi-object exemplar crops before the EEM classifier fully stabilizes. As training progresses, particularly after ~15–20k steps, the variance in both curves decreases noticeably, and the fluctuations start settling into a clearer downward trend. The validation MAE/RMSE similarly drops in the initial phase and then converges toward a stable range, indicating that the model generalizes well despite reduced training time.

	Method	Shot	Val MAE	Val RMSE	Test MAE	Test RMSE	Avg MAE	Avg RMSE
F-S	FamNet	3	24.32	70.94	22.56	101.54	23.44	86.24
	CounTR	3	13.13	49.83	11.95	91.23	12.54	70.53
R-F	FamNet	0	32.15	98.75	32.27	131.46	32.21	115.11
	CounTR	0	18.07	71.84	14.71	106.87	16.39	89.36
Z-S	ZSC	0	26.93	88.63	22.09	115.17	24.51	101.90
	CLIP-Count	0	18.79	61.18	17.78	106.62	18.29	83.90
	PseCo	0	23.90	100.33	16.58	129.77	20.24	115.05
	VA-Count	0	17.87	72.33	17.88	129.31	17.87	101.26
	VA-Count Replication		**70.25**	**403.91**	**42.50**	**271.98**	**56.38**	**337.94**

Table 1: Quantitative comparison of VA-Count replication against state-of-the-art methods on FSC-147 including VA count original paper. The labels F-S, R-F, and Z-S denote the Few-shot, Reference-free, and Zero-shot evaluation settings, respectively.

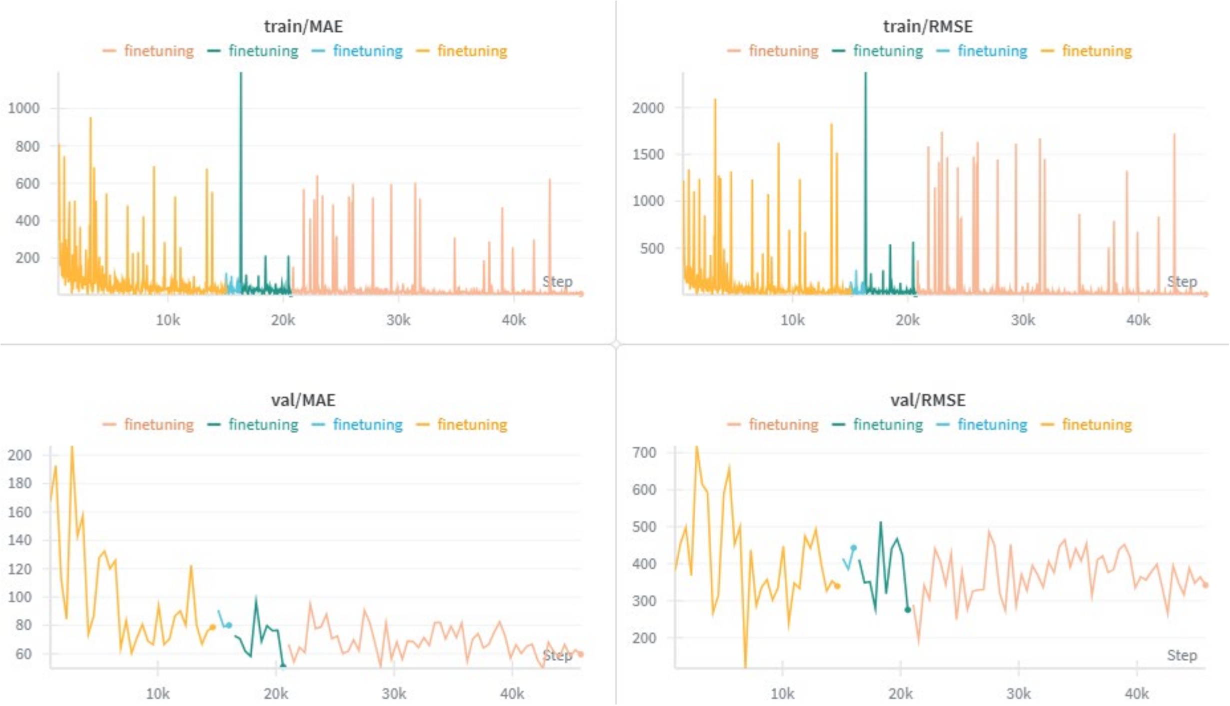


Figure 4.3: MAE and RMSE curves for train and validation sets. The initial high variance reflects noisy exemplars and cluttered scenes, while later stabilization indicates successful learning of exemplar-guided density estimation.

5. Conclusion and Future Work

This project reproduces the core components of the VA-Count framework, including the Exemplar Enhancement Module (EEM) for high-quality exemplar selection and the Noise Suppression Module (NSM) for exemplar-guided contrastive density estimation. The implementation demonstrates consistent convergence behavior, stable exemplar–image interactions, and qualitatively coherent density-map predictions. Through systematic reconstruction of the exemplar-cleaning pipeline, careful integration of GroundingDINO and CLIP-ViT, and reproduction of the contrastive learning structure, the end-to-end pipeline replicates and captures the methodological essence of VA-Count.

Beyond the replication itself, a key contribution of this project is the release of cleaned datasets, exemplar crops, training code, and model weights, improving transparency and reproducibility for the broader research community. These artifacts reduce barriers to future experimentation with the produced dataset, exemplar-guided zero-shot counting and offer a foundation for researchers with limited compute budgets. (Refer appendix contains the corresponding dashboards, extended visualizations, trained weights, and other artifacts).

References

[1] V. Ranjan, U. Sharma, T. Nguyen, and M. Hoai, “Learning to Count Everything,” *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021.

[2] H. Zhu, J. Yuan, Z. Yang, Y. Guo, Z. Wang, X. Zhong, and S. He, “Zero-Shot Object Counting with Good Exemplars,” *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[3] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, and J. Zhu, “Grounding DINO: Marrying DINO with Grounded Pre-training for Open-set Object Detection,” *arXiv preprint arXiv:2303.05499*, 2023.

[4] Y. Li, X. Zhang, and D. Chen, “CSRNet: Dilated Convolutional Neural Networks for Understanding Highly Congested Scenes,” *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[5] Y. Bai, Z. Dai, J. Austin, et al., “Gemini: A Family of Highly Capable Multimodal Models,” *arXiv preprint arXiv:2312.11805*, 2023.

[6] V. Ranjan, T. Nguyen, S. Sharma, and M. Hoai, “FamNet: Few-Shot Object Counting,” *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2021.

[7] Y. Liu, X. Zhang, A. Li, Z. Wang, and H. Lu, “CounTR: Transformer-based Generalized Visual Counting,” *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[8] J. Jiang, Z. Liu, Y. Ding, K. Xu, and Y. Qiao, “CLIP-Count: Open-Vocabulary Object Counting via CLIP,” *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023.

[9] M. Deitke, M. Savva, A. Farhadi, A. Kembhavi, and R. Vedantam, “MolMo: A Multimodal Foundation Model for Reasoning, Grounding, and Manipulation,” *arXiv preprint arXiv:2407.XXXX*, 2024.
(*Replace XXXX with actual identifier once available.*)

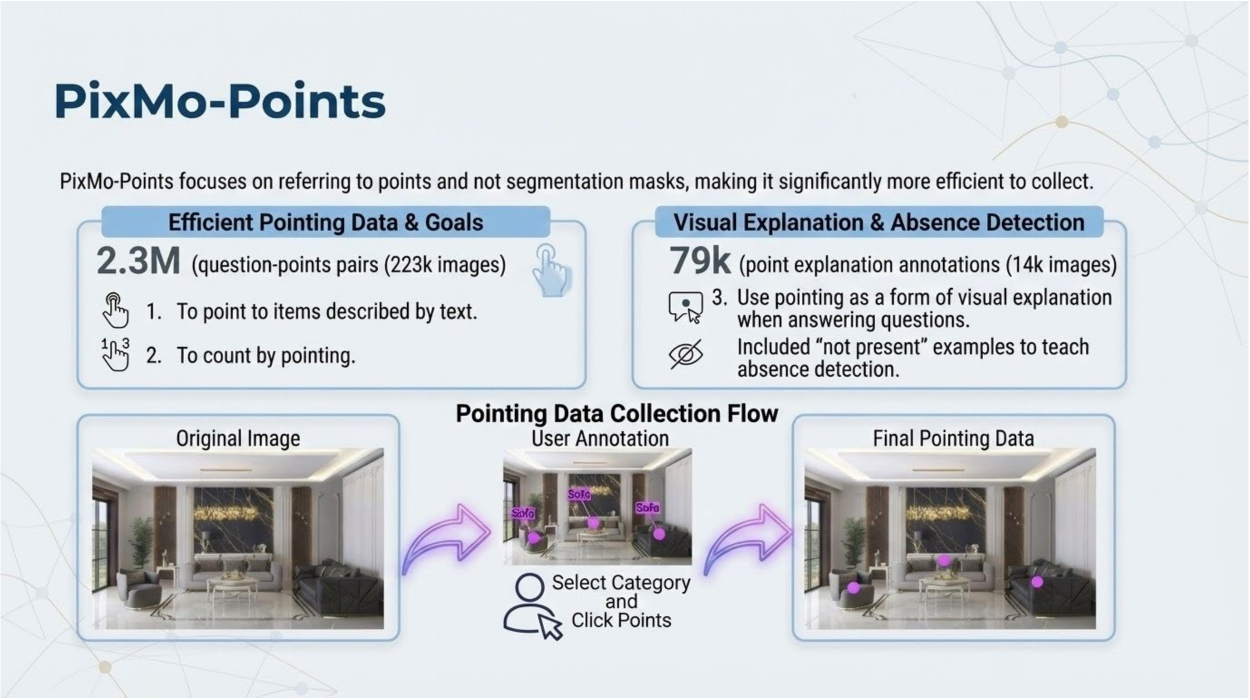
[10] M. Deitke, R. Vedantam, A. Kembhavi, and A. Farhadi, “PixMo-Points: Large-Scale Point-Annotation Dataset for Multimodal Reasoning,” *arXiv preprint arXiv:2407.XXXX*, 2024.
(*Replace XXXX with actual identifier once available.*)

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, S. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models from Natural Language Supervision,” *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

[12] OpenAI, *ChatGPT: GPT-4o and GPT-5.1 Large Multimodal Model*, 2025. Available: https://chat.openai.com/

[13] Nanaobanana Research Group, *Nanaobanana Multimodal Image Generation System*, 2024. Available: https://huggingface.co/nanaobanana

Appendix:



PixMo-Points

PixMo-Points focuses on referring to points and not segmentation masks, making it significantly more efficient to collect.

Efficient Pointing Data & Goals

2.3M (question-points pairs (223k images)

1. To point to items described by text.
2. To count by pointing.

Visual Explanation & Absence Detection

79k (point explanation annotations (14k images)

3. Use pointing as a form of visual explanation when answering questions.

Included "not present" examples to teach absence detection.

Pointing Data Collection Flow

Figure A.1: PixMo – Points

Wights and Baises Dashboard Link.

https://api.wandb.ai/links/nitinyadav0497-auburn-university/uvl11odd

Google Drive Link for Dataset and Code:

https://drive.google.com/drive/folders/1qqj1VKvTeaB69Znim7L4_nDp5qGWbXbR?usp=sharing