VQCounter: Designing Visual Prompt Queue for Accurate Open-World Counting

Fanfan Ye[1,2], **Yiqi Fan**[2], **Qiaoyong Zhong**[2*], **Shicai Yang**[2], **Di Xie**[2], **Jie Song**[1*], **Mingli Song**[1]
[1]Zhejiang University
[2]Hikvision Research Institute
{yeff, sjie, brooksong}@zju.edu.cn, {fanyiqi5, zhongqiaoyong, yangshicai, xiedi}@hikvision.com

Abstract

Class-agnostic counting enables enumerating arbitrary object classes beyond those seen during training. Recent studies attempted to exploit the potential of visual foundation models such as GroundingDINO. Despite the considerable progress, we observe certain shortcomings, including the limited diversity of visual prompts and suboptimal training regimen. To address these issues, we introduce VQCounter, which incorporates a visual prompt queue mechanism designed to enrich the diversity of visual prompts. A random modality switching strategy is proposed during training to strengthen both textual and visual modalities. Besides, in light of weak point supervision, a Voronoi diagram-based cost (VoronoiCost) is designed to improve Hungarian matching, leading to more stable and faster convergence. Building upon the Voronoi diagram, we also propose a novel set of more stringent evaluation metrics, which take point localization into account. Extensive experiments on the FSC-147 and CARPK datasets demonstrate that VQCounter achieves state-of-the-art performance in both zero-shot and few-shot settings, significantly outperforming existing methods across nearly all evaluations.

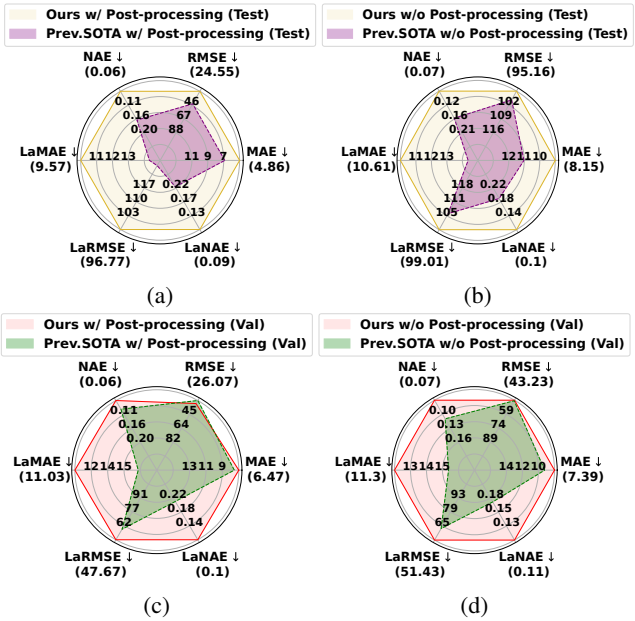


Figure 1: Performance comparison of VQCounter and the previous state-of-the-art method CountGD on FSC-147 (validation and test sets). Evaluations are based on six metrics: three standard metrics (MAE, RMSE, NAE) and their localization-aware counterparts (LaMAE, LaRMSE, LaNAE), both with and without post-processing. Lower values indicate better performance. In the radar chart, lower values are positioned at the periphery, while higher values are at the center.

1 Introduction

Object counting is an essential task with broad applications across agriculture, biomedicine, industry and so on. Historically, research has concentrated on class-specific counting, targeting categories such as human [Babu Sam *et al.*, 2022; Abousamra *et al.*, 2021; Li *et al.*, 2018; Liang *et al.*, 2022], animals [Arteta *et al.*, 2016; Zhu *et al.*, 2021; Sun *et al.*, 2023; Jia *et al.*, 2023], cells [Xie *et al.*, 2018; Guo *et al.*, 2019; Zheng *et al.*, 2024], and vehicles [Mundhenk *et al.*, 2016; Amato *et al.*, 2019]. Recently, attention has shifted towards class-agnostic counting (CAC) [Lu *et al.*, 2019], which empowers a model to count objects of arbitrary classes not encountered during training. Within the CAC paradigm, low-shot counting methods have gained considerable prominence, encompassing few-shot counting using visual exemplars [Liu *et al.*, 2022; Shi *et al.*, 2022; You *et al.*, 2023], zero-shot counting based on textual descriptions [Xu *et al.*, 2023; Ranjan and Nguyen, 2022; Amini-Naieni *et al.*, 2023; Kang *et al.*, 2024], and hybrid approaches that fuse both modalities [Amini-Naieni *et al.*, 2024; Dai *et al.*, 2024; Pelhan *et al.*, 2024b; Pelhan *et al.*, 2024a; Mondal *et al.*, 2024]. Advancements in large-scale, pre-trained vision-language models (VLMs) have further propelled CAC's state-of-the-art performance. CountGD [Amini-Naieni *et al.*, 2024], a leading method in both few-shot and zero-shot benchmarks, leverages the pre-trained vision-language model GroundingDINO [Liu *et al.*, 2025], an open-world text-specific object detector. By integrating visual exemplar features with textual tokens,

*Corresponding authors.

CountGD transforms GroundingDINO into an open-world counter, allowing users to specify objects for counting via text, visual exemplars, or a combination of both.

However, despite its exceptional performance, CountGD does not fully exploit GroundingDINO's potential, revealing avenues for enhancement. In particular, we identify four major limitations in CountGD. *(1) Insufficient Diversity of Visual Prompts.* CountGD heavily relies on the dataset for visual exemplars. For example, FSC-147 [Ranjan *et al.*, 2021] provides only three visual exemplars per image, which inadequately captures intra-class variations such as posture, color, and size. *(2) Simplified Label Matching.* Most counting datasets offer only point annotations, preventing the calculation of IoUCost for Hungarian matching [Kuhn, 1955] between ground truth and predictions. As a result, CountGD omits this critical component, leading to diminished convergence efficiency and overall performance. *(3) Monolithic Training Strategy.* CountGD exclusively focuses on the branch that combines text and visual exemplars during training, neglecting both purely textual and purely visual prompts. This approach limits performance when uni-modal prompts are used and undermines the full potential of multi-modal fusion. *(4) Inflated Counting Error Metrics.* Metrics like Mean Absolute Error (MAE) fail to accurately reflect localization errors. This issue is amplified in CountGD, where adaptive cropping reduces counting errors by balancing the numerous false positives and false negatives.

To overcome these limitations, in this paper we present VQCounter (**V**isual prompt **Q**ueue-based Counter), a refined yet highly effective open-world counting framework. As depicted in Figure 2, VQCounter addresses the aforementioned challenges and integrates novel localization-aware metrics to rigorously evaluate localization errors. *(1) Enhanced Visual Prompt Diversity.* To harness GroundingDINO's capacity to generate precise bounding boxes from point annotations, we introduce a dynamic queue that stores detection boxes throughout training, thereby augmenting the sparse visual exemplars within the dataset. Drawing inspiration from the Least Frequently Used (LFU) algorithm [Lee *et al.*, 2001], we implement a Most Frequently Used (MFU) queue management strategy. This technique prioritizes frequently used visual prompts for dequeuing while retaining less-utilized prompts for longer periods, thus bolstering the model's ability to accommodate intra-class variations. *(2) Improved Label Matching with VoronoiCost.* We introduce VoronoiCost, grounded in the Voronoi diagram [Aurenhammer and Klein, 2000], as a superior alternative to IoUCost for Hungarian matching. Unlike IoUCost, VoronoiCost is adept at handling point annotations, thereby enhancing both training efficiency and overall performance. *(3) Modality Switching Training Strategy.* To remedy the single-branch limitation, VQCounter adopts a tripartite training approach, incorporating three distinct prompt branches: text-only, visual exemplars-only, and a combined text and visual exemplars branch. By strengthening each modality independently, we enhance the efficacy of multi-modal prompt fusion. *(4) Localization-aware Metrics:* We devise novel metrics inspired by the Voronoi diagram, to assess localization errors in counting predictions. These metrics rely exclusively on ground truth data, ensuring a higher degree of objectivity compared to traditional methods.

We validate VQCounter through comprehensive experiments on two well-established datasets, namely FSC-147 [Ranjan *et al.*, 2021] and CARPK [Hsieh *et al.*, 2017]. The results clearly illustrate that VQCounter substantially surpasses existing methods in both zero-shot and few-shot benchmarks, as depicted in Figure 1.

2 Related Work

2.1 Existing Counting Tasks

Object counting has been extensively explored, primarily focusing on class-specific counting across various scenarios, including human [Abousamra *et al.*, 2021; Idrees *et al.*, 2018; Lian *et al.*, 2019; Song *et al.*, 2021], cells [Xie *et al.*, 2018; Guo *et al.*, 2019; Zheng *et al.*, 2024], animals [Arteta *et al.*, 2016], and polyps [Zavrtanik *et al.*, 2020]. While these methods achieve high accuracy for predefined categories, they often fail to count unseen classes during testing. In contrast, class-agnostic object counting approaches [Lu *et al.*, 2019; Zhu *et al.*, 2025; Liu *et al.*, 2022; Xu *et al.*, 2023] offer flexibility by accommodating arbitrary entities with or without prompts. Since the introduction of the FSC-147 benchmark [Ranjan *et al.*, 2021], significant progress has been made in zero-shot [Xu *et al.*, 2023; Ranjan and Nguyen, 2022; Amini-Naieni *et al.*, 2023; Jiang *et al.*, 2023] and few-shot [Lu *et al.*, 2019; Pelhan *et al.*, 2024b; Pelhan *et al.*, 2024a] counting, enhancing performance to unprecedented levels. Additionally, the class-agnostic paradigm has spurred novel tasks such as Referring Expression Counting [Dai *et al.*, 2024], Training-free Counting [Mondal *et al.*, 2024], and Unified Textual-Visual Prompts Counting [Amini-Naieni *et al.*, 2024], broadening the versatility of counting methodologies. In this paper, we adopt the class-agnostic framework, focusing on a textual-visual unified benchmark akin to CountGD [Amini-Naieni *et al.*, 2024], and improve performance by addressing previously unreported key issues.

2.2 Counting with VLMs

The advent of pre-trained vision-language models has significantly enhanced counting task performance. Zero-shot counting was first introduced by ZSC [Xu *et al.*, 2023], utilizing class names and CLIP's [Radford *et al.*, 2021] text encoder to generate prototypes for interaction with image feature maps. Subsequent models like CLIP-Counter [Jiang *et al.*, 2023] and VLCount [Kang *et al.*, 2024] refine the integration between text and image embeddings. Omini-Count [Mondal *et al.*, 2024] leverages the Segment Anything Model (SAM) [Kirillov *et al.*, 2023] alongside semantic and geometric priors for improved accuracy, while PseCo [Huang *et al.*, 2024] extends CLIP and SAM through a multi-task framework for precise instance segmentation. Recent advancements using GroundingDINO [Liu *et al.*, 2025], a state-of-the-art open-world detector, have set new benchmarks. VA-Count [Zhu *et al.*, 2025] enhances exemplar selection with exemplar enhancement and noise suppression, and GroundingREC [Dai *et al.*, 2024] introduces advanced feature fusion techniques and a fine-grained counting benchmark. CountGD [Amini-Naieni *et al.*, 2024] further improves

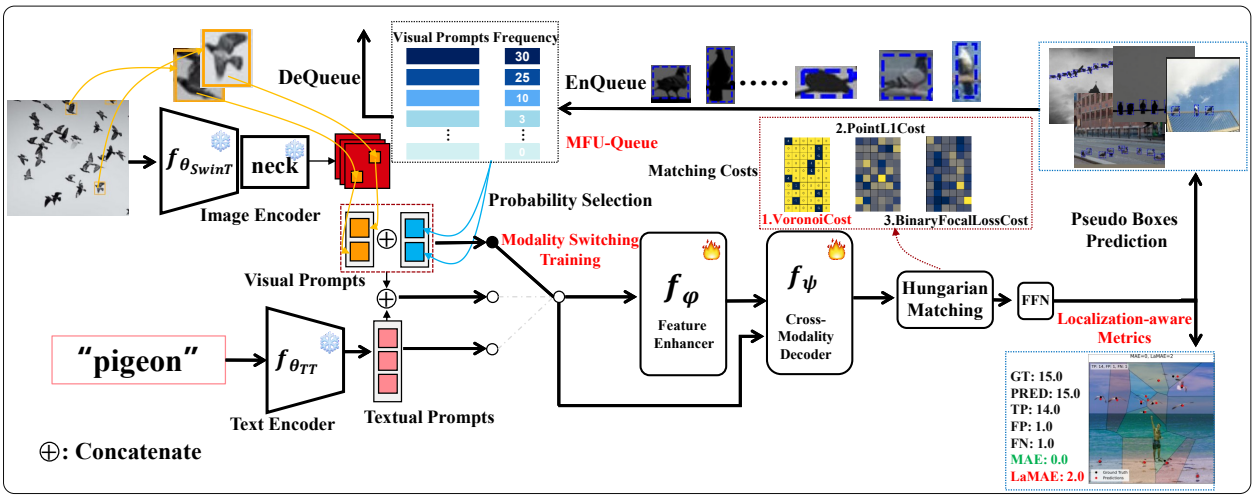


Figure 2: Pipeline of the proposed VQCounter framework. Building on CountGD, VQCounter introduces four key enhancements: **MFU-Queue** for input optimization, **VoronoiCost** as the primary algorithmic component, **Modality Switching Training** for the training methodology, and novel **Localization-aware Metrics**. In the diagram, these enhancements are highlighted in **bold red font**.

these methods by incorporating an image-text fusion module, enabling the model to process both textual and visual prompts effectively. Inspired by CountGD, our approach leverages GroundingDINO more efficiently, significantly increasing visual prompt diversity and enhancing class-agnostic counting performance.

2.3 Counting Metrics

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are prevalent in counting tasks due to their simplicity and computational efficiency. However, these metrics do not account for localization errors, as false positives and false negatives can cancel each other out, potentially misrepresenting model performance. To mitigate this issue, REC [Dai *et al.*, 2024] employs a manually defined threshold combined with Hungarian matching to compute true positives, false positives, and false negatives. Additionally, Song et al. [Song *et al.*, 2021] introduced the nAP metric, which utilizes a sophisticated sorting and filtering mechanism, while Ciampi et al. [Ciampi *et al.*, 2024] proposed the Mosaic Test to estimate coarse localization errors by splicing images. In this work, we introduce novel localization-aware metrics based on the Voronoi diagram [Aurenhammer and Klein, 2000], which accurately quantifies localization errors with enhanced interpretability and independence from predictions.

3 Method

In this section, we first briefly review the general practice in turning the pre-trained VLMs, *e.g.*, GroundingDINO [Liu *et al.*, 2025], into an open-word counter. Then, we elaborate on the proposed VQCounter framework and its key components. After that, novel localization-aware metrics are presented.

3.1 Preliminaries

Review of Counting with VLMs

We briefly review counting with pre-trained vision-language models, focusing on CountGD [Amini-Naieni *et al.*, 2024] as a representative of recent state-of-the-art approaches. In these methods, the target object is specified by visual exemplars (*e.g.*, bounding boxes $\mathbf{B} = \{b_1, \ldots, b_n\}$) or textual descriptions t. CAC aims to count entities of arbitrary classes, $\hat{y} = f(\mathbf{X}, \mathbf{B}, t)$, where $\mathbf{X}$ is the input image and $\hat{y}$ is the predicted count. CountGD consists of an Image Encoder ($f_{\theta_{SwinT}}$), a Text Encoder ($f_{\theta_{TT}}$), a Query Selection Module ($Select$), a Feature Enhancement Module (f_φ), and a Cross-Modality Decoder (f_ψ). After extracting visual prompts $\mathbf{P}_{visual} = RoIAlign\left(f_{\theta_{SwinT}}(\mathbf{X}), \mathbf{B}\right)$ and textual prompts $\mathbf{P}_{textual} = f_{\theta_{TT}}(t)$, the counting process is then formulated as follows:

$$(\mathbf{z}_{\mathbf{v},\mathbf{t}}, \mathbf{z}_{\mathbf{I}}) = f_\varphi\left(f_{\theta_{SwinT}}(\mathbf{X}), [\mathbf{P}_{visual}, \mathbf{P}_{text}]\right), \quad (1)$$

$$\hat{\mathbf{Y}} = Sigmoid\left(f_\psi\left(\mathbf{z}_{\mathbf{I}}, \mathbf{z}_{\mathbf{v},\mathbf{t}}, Select(\mathbf{z}_{\mathbf{I}}, \mathbf{z}_{\mathbf{I}}{\mathbf{z}_{\mathbf{v},\mathbf{t}}}^T, k)\right){\mathbf{z}_{\mathbf{v},\mathbf{t}}}^T\right), \quad (2)$$

where $\mathbf{z}_{\mathbf{v},\mathbf{t}}$ are the fused visual-textual features, $\mathbf{z}_{\mathbf{I}}$ are the image features, k is the number of selected image tokens, $\hat{\mathbf{Y}}$ are the final similarity scores, thresholded by the confidence threshold σ and counted to estimate the object count $\hat{y}$ during inference.

Discovery

During the replication of CountGD, we discover a surprising phenomenon that proves highly beneficial for low-shot CAC tasks, which has not been previously reported. We find that even when trained solely with point annotations, the model retains the ability to generate precise bounding boxes. As shown in the upper right corner of Figure 2, the predicted bounding boxes are accurate, and can be utilized as supplementary visual prompts. Their effectiveness is fully demonstrated in the experiments. Consequently, visual exemplars will be expanded to $\mathbf{B} \cup \hat{\mathbf{B}}$, where $\hat{\mathbf{B}} = \left\{\hat{b}_1, \ldots, \hat{b}_m\right\}$, $\hat{b}_i$ denotes a pseudo bounding box.

Algorithm 1: MFU-Queue Algorithm

Data: capacity E, insert batch size n, access batch size m, incoming vectors $\mathcal{V}_{\text{in}} = \{v_1, \ldots, v_n\}$
Result: MFU Queue $\mathbf{Q}_i$ with frequencies $f_i(v)$

Initialize $\mathbf{Q}_i \leftarrow \emptyset$ and $f_i(v) \leftarrow 0$ for all v;
while *training* **do**
  Insertion Phase;
  foreach $v_{new} \in \mathcal{V}_{in}$ **do**
    if $|\mathbf{Q}_i| = E$ **then**
      $v_{\text{remove}} \leftarrow \arg\max_{v \in \mathbf{Q}_i} f_i(v)$;
      $\mathbf{Q}_i \leftarrow \mathbf{Q}_i \setminus \{v_{\text{remove}}\}$;
    $\mathbf{Q}_i \leftarrow \mathbf{Q}_i \cup \{v_{\text{new}}\}$;
    $f_i(v_{\text{new}}) \leftarrow 0$;
  Access Phase;
  Select m vectors $\mathcal{V}_{\text{access}}$ from $\mathbf{Q}_i$ with probability $P(v) = \frac{1/(f_i(v)+\epsilon)}{\sum_{v' \in \mathbf{Q}_i} 1/(f_i(v')+\epsilon)}$;
  foreach $v \in \mathcal{V}_{access}$ **do**
    $f_i(v) \leftarrow f_i(v) + 1$;

3.2 VQCounter

In this part, we start to introduce the proposed counting framework VQCounter, a **V**isual prompt **Q**ueue-based Counter. In VQCounter we propose four improvements, namely MFU-Queue, VoronoiCost, Modality Switching Training, and Localization-aware Metrics, to improve the algorithm from a global perspective. Each part will be elaborated on in detail as follows.

MFU-Queue

Based on the description in Section 3.1, the pseudo yet accurate bounding boxes produced by the model could significantly enrich the diversity of visual exemplars. As shown in the upper right corner of Figure 2, even in the first training epoch, the model can detect several precise bounding boxes, and as training progresses, the recall of the detected boxes will gradually increase (more visualizations in the Appendix). These boxes can serve as a supplement to visual examples to enrich the diversity of visual prompts. To store these additional visual prompts, we first construct FIFO (First in, First Out) queues $\mathbf{Q} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_{\mathrm{K}}\}, \mathbf{Q}_i \in \mathbf{R}^{E \times d}$, where K denotes the number of classes in the training dataset and the capacity of each queue is set as a constant value E. In each training iteration, the visual prompts are drawn randomly from the ground truth boxes in the dataset and the prompt queues $\mathbf{Q}$.

Although the baseline queue helps to improve the diversity and effectiveness of visual prompts, the strategy for prompt entry, exit and sampling can be further optimized. Inspired by the Least Frequently Used (LFU) algorithm [Lee *et al.*, 2001], a famous cache algorithm for memory management within a computer, we propose MFU-Queue, a **M**ost **F**requently **U**sed priority dequeue **Queue**. As its name suggests, MFU-Queue tracks how often each prompt is accessed. When the queue is full, the most frequently accessed prompt is removed first. Simultaneously, within the queue, we employ a probabilistic accessing strategy that prioritizes prompts with fewer prior selections. During training, the enqueue and dequeue operations and access to visual prompts are performed alternately. Therefore, MFU-Queue significantly enhances the richness of visual prompts and demonstrates greater robustness to variations in object size and background within the same category. The MFU-Queue training procedure is summarized in Algorithm 1.

VoronoiCost

MFU-Queue optimizes the algorithm's input, while VoronoiCost, discussed in this section, enhances the algorithm's key components. One of the key components in DETR-like detectors is Hungarian matching between N object queries and M ground truth (GT) boxes. Its goal is to seek a matching function σ that assigns each GT box to a predicted box, that is $\sigma : \{1, 2, \ldots, M\} \rightarrow \{1, 2, \ldots, N\}$. Here, $\sigma(i)$ outputs the index of the prediction matched to the i-th GT box. The Hungarian algorithm is utilized to find the optimal matching σ^* that minimizes the total cost:

$$\sigma^* = \arg\min_{\sigma} \sum_{i=1}^{M} \mathbf{C}_{i,\sigma(i)}, \tag{3}$$

$\mathbf{C} \in R^{M \times N}$ is a cost matrix, where each element $\mathbf{C}_{i,j}$ represents the cost of matching the i-th ground truth box with the j-th predicted box. In DETR-like detectors, this cost typically combines classification cost, bounding box coordinates cost, and GIoU cost (α, β, and γ are coefficients):

$$\mathbf{C} = \alpha \mathbf{C}_{cls} + \beta \mathbf{C}_{coord} + \underline{\gamma \mathbf{C}_{GIoU}}. \tag{4}$$

Object counting datasets typically provide point annotations, where only the center point of each object is given, and the width and height are unknown. Consequently, in existing methods, $\mathbf{C}_{coord}$ only considers point coordinates, and the $\gamma \mathbf{C}_{GIoU}$ marked with underline is discarded, leading to suboptimal performance and efficiency. In this paper, a novel VoronoiCost (denoted as $\mathbf{C}_{Voron}$) is proposed to make up for the lack of $\mathbf{C}_{GIoU}$ when only point annotations are available in counting datasets.

VoronoiCost is constructed upon the Voronoi diagram [Aurenhammer and Klein, 2000], which is a partition of a plane into regions close to each of a given set of points. Each region corresponds to a point in the given point set, and the distance from any point within a region to its corresponding point is less than the distance to any other points.

Formally, given a set of GT points $\mathcal{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_M\}$, we first construct their Voronoi Polygons as:

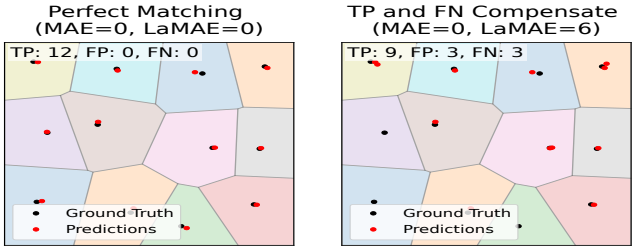
$$V(\mathbf{g}_i) = \left\{ \mathbf{x} \in \mathbb{R}^2 \mid d(\mathbf{x}, \mathbf{g}_i) < d(\mathbf{x}, \mathbf{g}_j) \, \forall \mathbf{g}_j \in \mathcal{G}, j \neq i \right\}, \tag{5}$$

where $d(\cdot)$ denotes euclidean distance. Then, given a set of predictions $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$, we can calculate $\mathbf{C}_{Voron}$ as follows:

$$\mathbf{C}_{Voron}(i,j) = c_{\text{sml}} \cdot \mathbb{I}(\mathbf{p}_j \in V(\mathbf{g}_i)) + c_{\text{lag}} \cdot \mathbb{I}(\mathbf{p}_j \notin V(\mathbf{g}_i)), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and c_{sml} and c_{lag} are constant values representing the matching cost when the predicted point $\mathbf{p}$ is within the Voronoi polygon area of the GT point $\mathbf{g}$. Eq. (4) now is turned as:

$$\mathbf{C} = \alpha \mathbf{C}_{cls} + \beta \mathbf{C}_{coord} + \gamma \mathbf{C}_{Voron}. \tag{7}$$



(a) (b)

Figure 3: Comparison of MAE and LaMAE. (a) Predictions perfectly match the ground truth, resulting in both MAE and LaMAE being 0. (b) False positives and false negatives cancel each other out, yielding a MAE of 0 but a LaMAE of 6, indicating the presence of localization errors.

Experimental results show the $\mathbf{C}_{Voron}$ boosts the performance and matching efficiency remarkably.

Modality Switching Training

In contrast to previous approaches that consider either uni-modal or multi-modal prompts in isolation, we propose a more generalized framework known as Modality Switching Training (MST). The key insight behind this design is that improving the performance of any individual modal prompts can enhance the effectiveness of multi-modal fused prompts. Specifically, VQCounter employs a gating controller that regulates the modality of prompts during training. Formally, the equation in Eq. (1) is rewritten as follows:

$$(\tilde{\mathbf{z}}, \mathbf{z}_{\mathbf{I}}) = f_{\varphi}\left(f_{\theta_{SwinT}}(\mathbf{X}), \tilde{\mathbf{P}}\right), \tag{8}$$

where $\tilde{\mathbf{P}} \in \{\mathbf{P}_{visual}, \mathbf{P}_{textual}, [\mathbf{P}_{visual}, \mathbf{P}_{textual}]\}$, and $\tilde{\mathbf{z}} \in \{\mathbf{z}_{\mathbf{v}}, \mathbf{z}_{\mathbf{t}}, \mathbf{z}_{\mathbf{v},\mathbf{t}}\}$. Meanwhile, the $\mathbf{z}_{\mathbf{v},\mathbf{t}}$ in the Eq.(2) will be replaced by $\tilde{\mathbf{z}}$ too. Building upon the MST strategy, the model's performance is reliably ensured for any uni-modal prompts, which significantly contributes to enhancing its performance in multi-modal prompts.

Localization-aware Metrics

To evaluate the algorithm's capabilities with greater precision and objectivity, this part introduces the novel localization-aware metrics. The motivation aligns with previous studies [Dai *et al.*, 2024; Song *et al.*, 2021; Ciampi *et al.*, 2024]. Common error metrics, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Normalized Absolute Error (NAE) fail to capture localization errors, potentially yielding accurate counts but imprecise localization. Specifically, false positives (FPs) and false negatives (FNs) may cancel each other out, leading to inflated performance metrics.

Unlike previous methods, the proposed localization-aware metrics are also built on the Voronoi diagram and offer greater objectivity and accuracy without relying on the information of predicted results. Before computing the true positives (TPs), false positives (FPs), and false negatives (FNs), we first define the number of predicted points within each Voronoi Polygon s_j as follows:

$$s_j = \left|\{\mathbf{p}_i \in \mathcal{P} \mid \mathbf{p}_i \in V(\mathbf{g}_j)\}\right|, \tag{9}$$

where $V(\mathbf{g}_j)$, $\mathcal{P}$, and range of i and j is the same as Eq. (5). Each Voronoi polygon containing at least one predicted point is counted as one TP, and any predicted points beyond the first are considered FPs. Each Voronoi polygon with no predicted points is counted as one FN, indicating a missed target.

Assume that the dataset size is B, and there are M_b GT points in the b-th image. Based on the counted TPs, FPs and FNs, we define a set of Localization-aware (La) metrics, including LaMAE, LaRMSE and LaNAE, as follows:

$$\begin{cases} \mathrm{TP_b} = \sum_{j=1}^{M_b} \mathbb{I}(s_j \geq 1) \\ \mathrm{FP_b} = \sum_{j=1}^{M_b} \max(s_j - 1, 0) \\ \mathrm{FN_b} = \sum_{j=1}^{M_b} \mathbb{I}(s_j = 0) = M_b - \mathrm{TP_b} \end{cases} \Rightarrow \begin{cases} \mathrm{LaMAE} = \frac{1}{B}\sum_{b=1}^{B}(FP_b + FN_b) \\ \mathrm{LaRMSE} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(FP_b + FN_b)^2}. \\ \mathrm{LaNAE} = \frac{1}{B}\sum_{b=1}^{B}\frac{FP_b + FN_b}{M_b} \end{cases} \tag{10}$$

These localization-aware metrics not only quantify counting errors but also assess localization inaccuracies. Taking MAE and LaMAE as an example, Figure 3 illustrates two typical scenarios. Both Figure 3a and Figure 3b exhibit MAE of 0. In Figure 3a, the absence of false positives and false negatives results in a LaMAE of 0. Conversely, Figure 3b includes FPs and FNs, leading to a LaMAE of 6. LaMAE is clearly more reasonable and accurate than MAE.

4 Experiments

In this section, we perform comprehensive experiments on two widely used datasets, *i.e.* FSC-147 [Ranjan *et al.*, 2021] and CARPK [Hsieh *et al.*, 2017], to validate the efficacy of our proposed VQCounter. To ensure a fair comparison, we adopt the same experimental setup as previous studies: VQCounter is trained using the FSC-147 training set and subsequently evaluated on the FSC-147 test set as well as the CARPK dataset without any fine-tuning.

4.1 Datasets & Metrics

FSC-147 consists of 6,135 images spanning 89 training, 29 validation, and 29 test classes, with each set containing mutually exclusive classes. Every image is annotated with at least three visual exemplars. **CARPK** includes 989 training and 459 test images of parking lots captured by overhead drones. Each image is annotated with at least two bounding boxes. Adopting the approach from [Liu *et al.*, 2022], we utilize these bounding boxes as visual exemplars and use the class name "car" for textual descriptions. More details about the two datasets and preprocessing are provided in the Appendix. **For evaluation metrics**, we employ both the standard metrics commonly used in counting tasks, namely MAE, RMSE, and NAE, as well as their localization-aware counterparts: LaMAE, LaRMSE, and LaNAE. These localization-aware metrics integrate counting and localization errors, offering a more comprehensive evaluation framework.

4.2 Implementation Details

For training, we use strategies similar to CountGD with the Adam optimizer and train for 30 epochs. Unlike CountGD, we fix $f_{\theta_{SwinT}}$ and $f_{\theta_{TT}}$ during training, and also stabilize

Method	Prompt Modality	Validation						Test					
		MAE ↓	RMSE ↓	NAE ↓	LaMAE ↓	LaRMSE ↓	LaNAE ↓	MAE ↓	RMSE ↓	NAE ↓	LaMAE ↓	LaRMSE ↓	LaNAE ↓
Patch-selection	Textual	26.93	88.63	-	-	-	-	22.09	115.17	-	-	-	-
CLIP-count	Textual	18.79	61.18	-	-	-	-	17.78	106.62	-	-	-	-
VLCounter	Textual	18.06	65.13	-	-	-	-	17.05	106.16	-	-	-	-
CounTX	Textual	17.1	65.61	-	-	-	-	15.88	106.29	-	-	-	-
DAVE_{prm}	Textual	15.48	52.57	-	-	-	-	14.9	103.42	-	-	-	-
GroundingREC	Textual	10.06	58.62	-	-	-	-	10.12	107.19	-	-	-	-
CountGD_{txt}	Textual	12.14	**47.51**	-	-	-	-	14.76	120.42	-	-	-	-
CountGD	Textual	12.21	69.45	0.13	18.79	80.22	0.18	15.02	131.74	0.14	18.57	133.56	0.18
VQCounter (Ours)	Textual	**8.72**	48.74	**0.07**	**12.52**	**57.54**	**0.11**	**6.84**	**84.04**	**0.06**	**9.79**	**92.51**	**0.09**
CounTR	Visual-I	13.13	49.83	-	-	-	-	11.95	91.23	-	-	-	-
LOCA	Visual-I	10.24	32.56	-	-	-	-	10.79	56.97	-	-	-	-
DAVE	Visual-I	8.91	**28.08**	-	-	-	-	8.66	32.36	-	-	-	-
CountGD	Visual-I	**8.13**	39.02	**0.09**	15.54	62.55	0.15	7.34	82.43	0.08	11.27	91.76	0.11
VQCounter (Ours)	Visual-I	8.69	39.22	0.1	**13.34**	**56.29**	**0.14**	**5.29**	**39.52**	**0.07**	**9.43**	**87.47**	**0.1**
CountGD	Visual-I & Textual	7.1	**26.07**	0.09	16.42	58.61	0.21	6.75	43.66	0.16	14.27	127.51	0.22
VQCounter (Ours)	Visual-I & Textual	**6.47**	30.15	**0.06**	**11.03**	**47.67**	**0.1**	**4.86**	**24.55**	**0.06**	**9.57**	**96.77**	**0.09**

Table 1: Comparison of state-of-the-art CAC methods on FSC-147 using textual-only, visual-only, and combined prompts. "Visual-I" denotes visual prompts from the current image. $_{txt}$ indicates retraining of the model using textual prompt only. VQCounter uses a unified model without retraining, offering greater practicality.

Method	Prompt Modality	Test		
		MAE (LaMAE) ↓	RMSE (LaRMSE) ↓	NAE (LaNAE) ↓
CLIP-count	Textual	11.96	16.61	-
CounTX	Textual	8.13	10.87	-
CountGD	Textual	3.83	5.41	-
VQCounter (Ours)	Textual	**2.74 (6.22)**	**3.77 (7.42)**	**0.04 (0.08)**
LOCA	Visual-I	9.97	12.51	-
CounTR	Visual-I	5.75	7.45	-
SAFECount	Visual-I	5.33	7.04	-
VQCounter (Ours)	Visual-I	**2.97 (6.36)**	**4.39 (7.66)**	**0.09 (0.13)**
CountGD	Visual-I & Textual	3.68	5.17	-
VQCounter (Ours)	Visual-I & Textual	**2.54 (6.22)**	**3.49 (7.45)**	**0.04 (0.08)**

Table 2: Comparison of state-of-the-art methods on CARPK using the same settings as Table 1. Values in parentheses indicate results evaluated using the localization-aware metrics.

the neck module, which adjusts feature dimensions, to ensure consistency in the MFU-Queue. Additionally, we adopt a non-parametric feature aggregation method from [Xu *et al.*, 2024; Ren, 2015], replacing a single 1×1 convolution layer with multi-level feature fusion. **During inference**, we use simple cropping instead of adaptive cropping, which introduces a significant number of false positives and false negatives. All other settings are consistent with CountGD. See the Appendix for more details.

4.3 Comparison with Other SOTA Methods

In this section, we evaluate the proposed VQCounter against other state-of-the-art methods using the widely recognized CAC datasets, FSC-147 and CARPK. Following CountGD, VQCounter is assessed with various modal prompts, *i.e.* textual-only, visual-only, and textual-visual combinations. For visual prompt, previous works focused on the interactive (Visual-I) setting, in which the visual prompts come from the current test image. Inspired by [Jiang *et al.*, 2025], we introduce an additional generic (Visual-G) setting, in which a fixed set of prompts are shared across the test set. As presented in Table 1 and 2, VQCounter significantly outperforms existing methods in nearly all metrics. On the FSC-147 benchmark, VQCounter outperforms the previous state-of-the-art method, CountGD, by reducing MAE by approximately 8% (from 7.1 to 6.47) on the validation set and by 28% (from 6.75 to 4.86) on the test set. To the best of our knowledge, this is the first method to achieve MAE values below 7 and 5 on the validation and test sets, respectively. Remarkably, in the zero-shot setting utilizing textual-only prompts, VQCounter surpasses CountGD by up to 50% (from 14.76 to 6.84), despite being trained exclusively on textual-based prompts, whereas our model employs a unified architecture. Furthermore, VQCounter demonstrates superior performance under the more stringent localization-aware metrics.

On the CARPK benchmark, we present results evaluated using both standard and localization-aware metrics (detailed in brackets). Consistently, VQCounter achieves state-of-the-art performance across all three prompt categories, significantly outperforming existing approaches.

4.4 Comparison to CountGD in More Aspects

For a detailed evaluation, we create a subset by selecting samples containing fewer than 900 targets for both validation and test sets. This subset fits well within the detection limit of the model, allowing for a single inference without the necessity of multi-cropping inference. Additionally, we investigate the impact of the two post-processing strategies, namely Adaptive Cropping and SAM TT-Norm proposed by [Amini-Naieni *et al.*, 2024]. Table 3 presents a comprehensive comparison between VQCounter and CountGD with (filled with gray) and without these post-processing strategies on both of the complete dataset and the newly formed subset.

Of the 156 metrics evaluated, VQCounter achieves superior results in 148 cases, with many metrics showing substantial improvements over CountGD. Similarly, in terms of the localization-aware metrics, VQCounter demonstrates overall superior performance compared to CountGD. In a few cases the performance decreases in the Visual-I setting. We hypothesize that MFU-Queue, while improving classes with significant intra-class variation, may slightly hurt classes with minor intra-class variation. However, this impact is negligible. Results in the Visual-G setting show that when visual prompts are sourced from different images, our method significantly outperforms CountGD, supporting our hypothesis.

4.5 Ablation Studies

We conduct some ablation studies to explore the proposed method systematically. For simplicity, we report results without the Adaptive Cropping and SAM TT-Norm [Amini-Naieni *et al.*, 2024] post-processing. As illustrated in Table 4,

Method	Prompt Modality	Split	Validation MAE ↓	RMSE ↓	NAE ↓	LaMAE ↓	LaRMSE ↓	LaNAE ↓	Test MAE ↓	RMSE ↓	NAE ↓	LaMAE ↓	LaRMSE ↓	LaNAE ↓
CountGD	T&V-I	all	8.69	43.89	0.11	15.83	62.77	0.17	10.92	99.58	0.16	14.45	106.17	0.19
		≤ 900	6.94	24.39	0.11	12.49	36.73	0.17	7.18	31.76	0.16	10.53	35.74	0.19
Ours		all	7.39 (-1.30)	43.23 (-0.66)	0.07 (-0.04)	11.3 (-4.53)	51.43 (-11.34)	0.11 (-0.06)	8.15 (-2.77)	95.16 (-4.42)	0.07 (-0.09)	10.61 (-3.84)	99.01 (-7.16)	0.1 (-0.09)
		≤ 900	5.35 (-1.59)	19.22 (-5.17)	0.07 (-0.04)	8.57 (-3.92)	25.93 (-10.80)	0.11 (-0.06)	4.41 (-2.77)	12.11 (-19.65)	0.07 (-0.09)	6.76 (-3.77)	16.07 (-19.67)	0.1 (-0.09)
CountGD	T	all	12.47	67.49	0.14	19.26	79.48	0.2	15.95	132.08	0.15	19.39	133.9	0.19
		≤ 900	8.9	28.93	0.14	14.74	41.26	0.19	10.81	35.16	0.15	14.23	38.76	0.19
Ours		all	8.29 (-4.18)	46.88 (-20.61)	0.07 (-0.07)	12.36 (-6.90)	56.2 (-23.28)	0.11 (-0.09)	8.82 (-7.13)	96.81 (-35.27)	0.08 (-0.07)	11.55 (-7.84)	104.28 (-29.62)	0.1 (-0.09)
		≤ 900	6.17 (-2.73)	24.47 (-4.46)	0.07 (-0.07)	9.49 (-5.25)	31.55 (-9.71)	0.11 (-0.08)	5.05 (-5.76)	17.11 (-18.05)	0.07 (-0.08)	7.55 (-6.68)	20.9 (-17.86)	0.1 (-0.09)
CountGD	V-I	all	9.33	51.76	0.1	16.23	67.47	0.16	10.05	96.14	0.1	13.81	104.55	0.14
		≤ 900	7.09	29.84	0.1	12.62	39.64	0.16	6.31	18.15	0.1	9.83	25.05	0.14
Ours		all	10.01 (+0.68)	52.81 (+1.05)	0.11 (+0.01)	13.93 (-2.30)	59.68 (-7.79)	0.15 (-0.01)	8.81 (-1.24)	95.74 (-0.40)	0.08 (-0.02)	11.42 (-2.39)	99.65 (-4.90)	0.11 (-0.03)
		≤ 900	7.68 (+0.59)	29.67 (-0.17)	0.11 (+0.01)	10.95 (-1.67)	35.09 (-4.55)	0.15 (-0.01)	5.06 (-1.25)	14.87 (-3.28)	0.08 (-0.02)	7.57 (-2.26)	18.6 (-6.45)	0.11 (-0.03)
CountGD	V-G	all	10.68	58.47	0.16	18.84	74.03	0.29	10.96	99.53	0.16	15.84	108.05	0.22
		≤ 900	8.4	40.96	0.16	15.21	49.31	0.28	7.23	31.63	0.16	11.85	37.5	0.22
Ours		all	9.73 (-0.95)	51.54 (-6.93)	0.11 (-0.05)	13.74 (-5.10)	58.51 (-15.52)	0.15 (-0.14)	8.84 (-2.12)	96.2 (-3.33)	0.08 (-0.08)	11.52 (-4.32)	101.52 (-6.53)	0.11 (-0.11)
		≤ 900	7.32 (-1.08)	26.93 (-14.03)	0.1 (-0.06)	10.77 (-4.44)	33.34 (-15.97)	0.15 (-0.13)	5.08 (-2.15)	15.25 (-16.38)	0.08 (-0.08)	7.61 (-4.24)	19.02 (-18.48)	0.11 (-0.11)
CountGD	T&V-I	≤ 900	6.22	21.2	0.09	13.3	36.61	0.2	6.2	39.28	0.16	10.72	43.34	0.22
Ours		≤ 900	5.1 (-1.12)	18.16 (-3.04)	0.06 (-0.03)	8.39 (-4.91)	25.28 (-11.33)	0.1 (-0.10)	3.97 (-2.23)	10.59 (-28.69)	0.06 (-0.10)	6.42 (-4.30)	15.11 (-28.23)	0.09 (-0.13)
CountGD	T	≤ 900	8.3	26.09	0.12	14.17	39.29	0.18	9.88	33.86	0.14	13.42	37.58	0.18
Ours		≤ 900	6.27 (-2.03)	24.96 (-1.13)	0.07 (-0.05)	9.42 (-4.75)	31.55 (-7.74)	0.11 (-0.07)	4.21 (-5.67)	11.36 (-22.50)	0.06 (-0.08)	6.81 (-6.61)	16.71 (-20.87)	0.09 (-0.09)
CountGD	V-I	≤ 900	6.67	28.39	0.09	12.22	38.55	0.15	4.98	14.02	0.07	8.61	22.11	0.11
Ours		≤ 900	7.14 (+0.47)	26.83 (-1.56)	0.1 (+0.01)	10.35 (-1.87)	32.55 (-6.00)	0.14 (-0.01)	4.09 (-0.89)	10.85 (-3.17)	0.07 (0.00)	6.78 (-1.83)	16.07 (-6.04)	0.1 (-0.01)
CountGD	V-G	all	10.72	64.51	0.1	17.02	79.03	0.16	8.23	87.36	0.14	12.29	96.32	0.18
		≤ 900	7.26	29.83	0.1	12.56	39.27	0.16	5.84	29.17	0.14	9.6	34	0.18
Ours		all	9.29 (-1.43)	41.91 (-22.60)	0.09 (-0.01)	13.58 (-3.44)	56.57 (-22.46)	0.13 (-0.03)	5.39 (-2.84)	38.64 (-48.72)	0.07 (-0.07)	9.81 (-2.48)	87.96 (-8.36)	0.1 (-0.08)
		≤ 900	7.19 (-0.07)	26.07 (-3.76)	0.09 (-0.01)	10.54 (-2.02)	32.58 (-6.69)	0.13 (-0.03)	4.27 (-1.57)	11.39 (-17.78)	0.07 (-0.07)	6.99 (-2.61)	16.51 (-17.49)	0.1 (-0.08)

Table 3: A comprehensive comparison between VQCounter and CountGD. *all* and ≤ 900 denote the full dataset and the subset with ≤ 900 objects per image, respectively. *T*, *V-I*, *V-G*, and *T & V-I* represent Textual-only, Visual-I only, Visual-G only, and the combined prompts. Gray background indicates results with post-processing. Improved performance is shown in red, otherwise in green. Note: check Table 1 for the rest results with post-processing on the full set.

Baseline Queue	MFU Queue	Voronoi Cost	Split	MAE ↓	RMSE ↓	NAE ↓
✗	✗	✗	Test	9.24	96.07	0.08
			Val	9.43	54.43	0.09
✗	✗	✓	Test	8.87 (-0.37)	104.14 (+8.07)	0.07 (-0.01)
			Val	8.36 (-1.07)	50.57 (-3.86)	0.08 (-0.01)
✓	✗	✗	Test	8.8 (-0.44)	96.55 (+0.48)	0.08 (0.00)
			Val	8.14 (-1.29)	47.3 (-7.13)	0.07 (-0.02)
✓	✗	✓	Test	8.41 (-0.83)	95.4 (-0.67)	0.08 (0.00)
			Val	7.77 (-1.66)	47.63 (-6.80)	0.08 (-0.01)
✗	✓	✓	Test	8.15 (-1.09)	95.16 (-0.91)	0.07 (-0.01)
			Val	7.39 (-2.04)	43.23 (-11.20)	0.07 (-0.02)

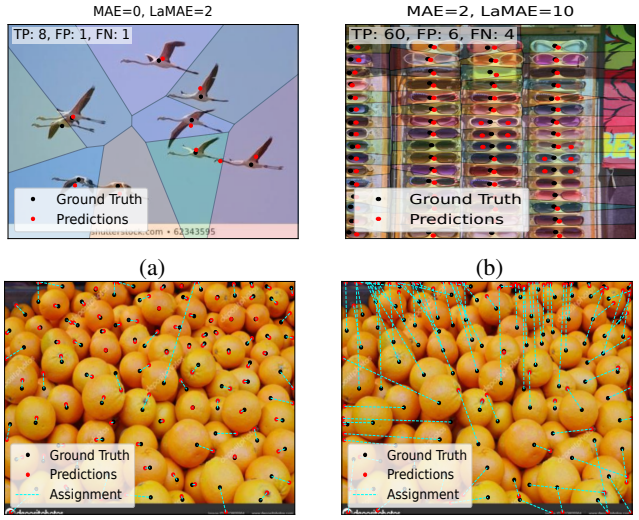
Baseline Queue	MFU Queue	Voronoi Cost	Split	LaMAE ↓	LaRMSE ↓	LaNAE ↓
✗	✗	✗	Test	13	105.31	0.12
			Val	15.75	67.91	0.14
✗	✗	✓	Test	11.93 (-1.07)	107.89 (+2.58)	0.12 (0.00)
			Val	12.74 (-3.01)	56.81 (-11.10)	0.16 (+0.02)
✓	✗	✗	Test	11.95 (-1.05)	102.98 (-2.33)	0.12 (0.00)
			Val	12.9 (-2.85)	54.45 (-13.46)	0.16 (+0.02)
✓	✗	✓	Test	11.02 (-1.98)	100.64 (-4.67)	0.11 (-0.01)
			Val	11.73 (-4.02)	55.25 (-12.66)	0.12 (-0.02)
✗	✓	✓	Test	10.61 (-2.39)	99.01 (-6.30)	0.1 (-0.02)
			Val	11.3 (-4.45)	51.43 (-16.48)	0.11 (-0.03)

Table 4: Ablation studies of VQCounter on FSC-147. Performance improvements are marked in red, while performance degradations are marked in green.

both VoronoiCost and the Baseline Queue independently improve the performance relative to the baseline method. Their combination yields additional performance gains. Furthermore, integrating the MFU strategy to optimize the Baseline Queue continues to augment the overall performance of the proposed method.

4.6 Visualization Analysis

To better understand the superiority of the localization-aware metrics and VoronoiCost, we visualize them in Figure 4. The difference between LaMAE and MAE is shown in Figure 4a and Figure 4b. LaMAE is stricter and can measure FPs and FNs. The Hungarian matching results between the predictions and the ground truth points in the first epoch are visualized in Figure 4c and Figure 4d. It can be seen that the label assignment converges faster after adding VoronoiCost. See more visualizations in the Appendix.



(c) with VoronoiCost (d) without VoronoiCost

Figure 4: Visualizations of the localization-aware metrics (a-b) and Hungarian matching results (c-d). Each cyan dashed line indicate a pair of matched prediction and GT point.

5 Conclusion

In summary, we present VQCounter, a superior open-world counting framework that enhances diversity of visual prompts with dynamic queues, utilizes VoronoiCost for better matching between predictions and GT points, and adopts the modality switching training approach for better collaborative learning of different modalities. Additionally, we introduce the Voronoi diagram-based metrics to accurately measure the localization errors. By exploiting vision-language models more effectively, we achieve state-of-the-art results on FSC-147 and CARPK in both zero-shot and few-shot settings, setting a new benchmark in class-agnostic counting.

References

[Abousamra *et al.*, 2021] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 872–881, 2021.

[Amato *et al.*, 2019] Giuseppe Amato, Luca Ciampi, Fabrizio Falchi, and Claudio Gennaro. Counting vehicles with deep learning in onboard uav imagery. In *2019 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2019.

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

[Amini-Naieni *et al.*, 2024] N. Amini-Naieni, T. Han, and A. Zisserman. Countgd: Multi-modal open-world counting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[Arteta *et al.*, 2016] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 483–498. Springer, 2016.

[Aurenhammer and Klein, 2000] Franz Aurenhammer and Rolf Klein. Voronoi diagrams. *Handbook of computational geometry*, 5(10):201–290, 2000.

[Babu Sam *et al.*, 2022] Deepak Babu Sam, Abhinav Agarwalla, Jimmy Joseph, Vishwanath A Sindagi, R Venkatesh Babu, and Vishal M Patel. Completely self-supervised crowd counting via distribution matching. In *ECCV*, pages 186–204. Springer, 2022.

[Ciampi *et al.*, 2024] Luca Ciampi, Nicola Messina, Matteo Pierucci, Giuseppe Amato, Marco Avvenuti, and Fabrizio Falchi. Mind the prompt: A novel benchmark for prompt-based class-agnostic counting. *arXiv preprint arXiv:2409.15953*, 2024.

[Dai *et al.*, 2024] Siyang Dai, Jun Liu, and Ngai-Man Cheung. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16985–16995, 2024.

[Guo *et al.*, 2019] Yue Guo, Jason Stein, Guorong Wu, and Ashok Krishnamurthy. Sau-net: A universal deep network for cell counting. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 299–306, 2019.

[Hsieh *et al.*, 2017] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017.

[Huang *et al.*, 2024] Zhizhong Huang, Mingliang Dai, Yi Zhang, Junping Zhang, and Hongming Shan. Point segment and count: A generalized framework for object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17067–17076, 2024.

[Idrees *et al.*, 2018] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–546, 2018.

[Jia *et al.*, 2023] Jieru Jia, Shuorui Zhang, and Qiuqi Ruan. Pcr: A large-scale benchmark for pig counting in real world. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 227–240. Springer, 2023.

[Jiang *et al.*, 2023] Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4535–4545, 2023.

[Jiang *et al.*, 2025] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pages 38–57. Springer, 2025.

[Kang *et al.*, 2024] 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, pages 2714–2722, 2024.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[Kuhn, 1955] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[Lee *et al.*, 2001] Donghee Lee, Jongmoo Choi, Jong-Hun Kim, Sam H Noh, Sang Lyul Min, Yookun Cho, and Chong Sang Kim. Lrfu: A spectrum of policies that subsumes the least recently used and least frequently used policies. *IEEE transactions on Computers*, 50(12):1352–1361, 2001.

[Li *et al.*, 2018] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[Lian *et al.*, 2019] Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao. Density map regression guided detection network for rgb-d crowd counting and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1821–1830, 2019.

[Liang *et al.*, 2022] Dingkang Liang, Wei Xu, and Xiang Bai. An end-to-end transformer model for crowd localization. In *European Conference on Computer Vision*, pages 38–54. Springer, 2022.

[Liu *et al.*, 2022] Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. Countr: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721*, 2022.

[Liu *et al.*, 2025] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025.

[Lu *et al.*, 2019] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 669–684. Springer, 2019.

[Mondal *et al.*, 2024] Anindya Mondal, Sauradip Nag, Xiatian Zhu, and Anjan Dutta. Omnicount: Multi-label object counting with semantic-geometric priors. *arXiv preprint arXiv:2403.05435*, 2024.

[Mundhenk *et al.*, 2016] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 785–800. Springer, 2016.

[Pelhan *et al.*, 2024a] Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A novel unified architecture for low-shot counting by detection and segmentation. *arXiv preprint arXiv:2409.18686*, 2024.

[Pelhan *et al.*, 2024b] Jer Pelhan, Vitjan Zavrtanik, Matej Kristan, et al. Dave-a detect-and-verify paradigm for low-shot counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23293–23302, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Ranjan and Nguyen, 2022] Viresh Ranjan and Minh Hoai Nguyen. Exemplar free class agnostic counting. In *Proceedings of the Asian Conference on Computer Vision*, pages 3121–3137, 2022.

[Ranjan *et al.*, 2021] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021.

[Ren, 2015] Shaoqing Ren. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[Shi *et al.*, 2022] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9529–9538, 2022.

[Song *et al.*, 2021] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3365–3374, 2021.

[Sun *et al.*, 2023] Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13791–13801, 2023.

[Xie *et al.*, 2018] Weidi Xie, J Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, 6(3):283–292, 2018.

[Xu *et al.*, 2023] 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*, pages 15548–15557, 2023.

[Xu *et al.*, 2024] Yifan Xu, Mengdan Zhang, Chaoyou Fu, Peixian Chen, Xiaoshan Yang, Ke Li, and Changsheng Xu. Multi-modal queried object detection in the wild. *Advances in Neural Information Processing Systems*, 36, 2024.

[You *et al.*, 2023] Zhiyuan You, Kai Yang, Wenhan Luo, Xin Lu, Lei Cui, and Xinyi Le. Few-shot object counting with similarity-aware feature enhancement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6315–6324, 2023.

[Zavrtanik *et al.*, 2020] Vitjan Zavrtanik, Martin Vodopivec, and Matej Kristan. A segmentation-based approach for polyp counting in the wild. *Engineering Applications of Artificial Intelligence*, 88:103399, 2020.

[Zheng *et al.*, 2024] Zixuan Zheng, Yilei Shi, Chunlei Li, Jingliang Hu, Xiao Xiang Zhu, and Lichao Mou. Rethinking cell counting methods: Decoupling counting and localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 418–426. Springer, 2024.

[Zhu *et al.*, 2021] Pengfei Zhu, Tao Peng, Dawei Du, Hongtao Yu, Libo Zhang, and Qinghua Hu. Graph regularized flow attention network for video animal counting from drones. *IEEE Transactions on Image Processing*, 30:5339–5351, 2021.

[Zhu *et al.*, 2025] Huilin Zhu, Jingling Yuan, Zhengwei Yang, Yu Guo, Zheng Wang, Xian Zhong, and Shengfeng He. Zero-shot object counting with good exemplars. In *European Conference on Computer Vision*, pages 368–385. Springer, 2025.