Decoupling What to Count and Where to See for Referring Expression Counting

Yuda Zou, Zijian Zhang, Yongchao Xu[†]

National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Hubei Luojia Laboratory, Wuhan University, Wuhan 430072, China
zouyuda@whu.edu.cn

Abstract

Referring Expression Counting (REC) extends class-level object counting to the fine-grained subclass-level, aiming to enumerate objects matching a textual expression that specifies both the class and distinguishing attribute. A fundamental challenge, however, has been overlooked: annotation points are typically placed on class-representative locations (*e.g.*, heads), forcing models to focus on class-level features while neglecting attribute information from other visual regions (*e.g.*, legs for "walking"). To address this, we propose *W2-Net*, a novel framework that explicitly decouples the problem into "what to count" and "where to see" via a dual-query mechanism. Specifically, alongside the standard what-to-count (w2c) queries that localize the object, we introduce dedicated where-to-see (w2s) queries. The w2s queries are guided to seek and extract features from attribute-specific visual regions, enabling precise subclass discrimination. Furthermore, we introduce Subclass Separable Matching (SSM), a novel matching strategy that incorporates a repulsive force to enhance inter-subclass separability during label assignment. *W2-Net* significantly outperforms the state-of-the-art on the REC-8K dataset, reducing counting error by 22.5% (validation) and 18.0% (test), and improving localization F1 by 7% and 8%, respectively.

Code — https://github.com/zouyuda220/W2-Net

Introduction

Object counting, a fundamental computer vision task, has diverse applications from crowd analysis (Han et al. 2023; Huang et al. 2024) and traffic management (Li et al. 2014) to cell biology (Chen et al. 2021). Early work in this domain, known as Class-Specific Counting (CSC), developed specialized models for predefined classes. Among these, Crowd Counting (Song et al. 2021) has been the most extensively studied, aiming to determine the number of people in an image. Crowd counting methods can generally be divided into density-based ones (Lin et al. 2024; Yang et al. 2025b) and localization-based ones (Song et al. 2021; Lin, Zhao, and Chan 2025; Liang, Xu, and Bai 2022). While effective, CSC methods are limited in flexibility, requiring a new model to be trained for each new class of interest. This spurred the development of Class-Agnostic Counting (CAC) (Ranjan et al. 2021), which generalizes to arbitrary classes using visual (Yang et al. 2025a) or text prompts (Shi et al. 2025; Wang et al. 2025a).

Despite these progress, both CSC and CAC models are limited to count objects at the class level, remaining obliv-

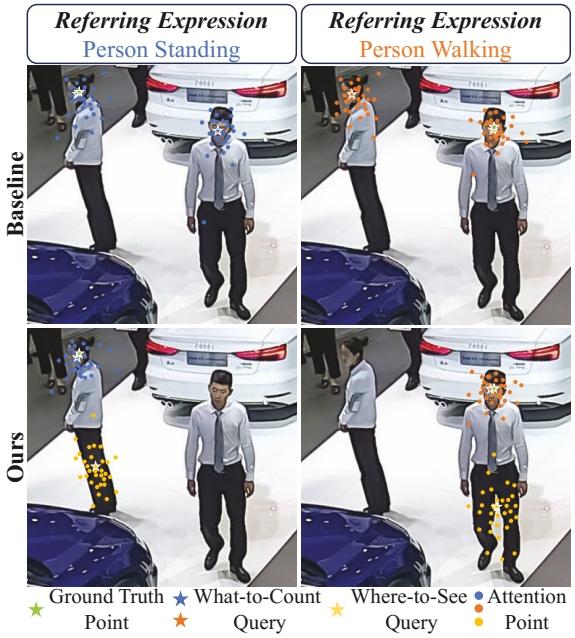

Figure 1: Illustration of the core challenge in REC we address. REC point annotations (green pentastar), placed on class-representative locations like heads, provide insufficient guidance for attribute-specific regions (*e.g.*, legs for "standing" or "walking"). This hinders the model from distinguishing fine-grained subclasses ("person standing" and "person walking"). Our *W2-Net* introduces a dedicated where-to-see (w2s) queries (yellow pentastar) that actively seek attribute-relevant visual cues. By fusing these features to the corresponding standard what-to-count (w2c), the model achieves precise subclass discrimination. The attention points visualize the attention focus of each query type. Best viewed by zooming in the electronic version.

[†]Corresponding author.

ious to the fine-grained attributes that define distinct subclasses within a class to bridge this gap. Referring Expression Counting (REC) (Dai, Liu, and Cheung 2024) has recently emerged, aiming to count objects of a specific subclass defined by a given textual referring expression (*e.g.*, "people walking" vs. "people standing"), which specifies both a class name and a distinguishing attribute. The key challenge in REC lies in distinguishing different subclasses that share the same class but differ in specific attributes (Triaridis et al. 2025). To alleviate this, the pioneering work, GroundingREC (Dai, Liu, and Cheung 2024), adopts a detection-then-count pipeline built on the powerful open-set detector GroundingDINO (Liu et al. 2024). It enhances the detector with global-local feature fusion and contrastive learning to differentiate the attributes. Subsequently, CAD-GD (Wang et al. 2025b) proposes a density-based module to guide the localization and counting process. While they have made promising strides, they overlook a critical limitation inherent in the REC's annotation scheme.

Following standard practice in object counting datasets (Zhang et al. 2016; Idrees et al. 2018; Sindagi, Yasarla, and Patel 2020), REC dataset (Dai, Liu, and Cheung 2024) uses a cost-effective point annotation format where each target object is marked with a single 2D point. These points are typically marked on class-representative locations like a person's head (see Fig. 1) without considering the visual regions that characterize the specific attributes. Used for supervision, these annotation points inherently force the model to over-emphasize class-level features extracted from their vicinity, while neglecting the crucial attribute-specific features located elsewhere (*e.g.*, legs for "walking"). As illustrated in Fig. 1, this lack of attribute-aware guidance makes it difficult for the model to distinguish between different subclasses (*e.g.*, "person walking" and "person standing") and ultimately impairs the counting performance.

To explicitly tackle this challenge, we propose a novel framework *W2-Net*. Following prior methods (Dai, Liu, and Cheung 2024; Wang et al. 2025b), *W2-Net* is built upon the powerful GroundingDINO detector (Liu et al. 2024), adapting it from box detection to point-based localization for the counting task. At the core of *W2-Net* is a novel W2 decoder, designed to decouple the problem into "what to count" and "where to see". Alongside the standard what-to-count (w2c) query that aims to localize the target object, we additionally introduce a dedicated where-to-see (w2s) query. Crucially, both queries are processed in parallel and iteratively refined through each decoder layer. While the w2c query learns to converge on the class-representative center (*e.g.*, a person's head) consistent with the supervised annotation point, the w2s query is specifically guided by the attribute portion in the referring expression. This guidance enables it to actively navigate the feature space and anchor itself to the visual region most indicative of the given attribute (*e.g.*, yellow pentastar and attention points on legs for "walking" and "standing" in Fig. 1). The features from these regions are then extracted by the w2s query and fused to enrich the w2c query, enabling improved subclass distinction and counting. Furthermore, we introduce Subclass Separable Matching (SSM) to alleviate the training instability arising from inter-subclass ambiguity. By incorporating a repulsive force into the matching cost, SSM actively penalizes assignments of w2c queries near non-target subclasses, thereby enforcing greater separability and ensuring more stable supervision.

Extensive experiments on the REC-8K benchmark (Dai, Liu, and Cheung 2024) demonstrate that our method significantly outperforms the state-of-the-art methods on both counting and localization. Specifically, our approach achieves a remarkable reduction in relative counting error by 22.5% on the validation set and 18.0% on the test set. Concurrently, it boosts the localization F1-score by 7% and 8%, respectively. Our contributions are three-fold:

- To the best of our knowledge, we are the first to identify and analyze the annotation issue in REC.
- We propose a novel framework, *W2-Net*, and a subclass separable matching, effectively enhancing the attribute perception and subclass discrimination.
- We set a new state-of-the-art on REC, significantly outperforming prior methods in counting and localization.

Related Work

The Evolution of Object Counting

The field of object counting (Pothiraj et al. 2025; Dumery et al. 2025) has evolved significantly, moving from constrained, class-specific counting to more versatile and flexible paradigms (Mondal et al. 2025; Guo, Gao, and Yuan 2025; Perez, Maji, and Sheldon 2024; He et al. 2024). Initial efforts focused on Class-Specific Counting (CSC), where models were tailored to enumerate instances of a single, pre-defined class, such as people (Han et al. 2023) or vehicles (Hsieh, Lin, and Hsu 2017). These methods, typically based on indirect density map regression (Sun et al. 2023) or direct object localization (Song et al. 2021; Liu et al. 2023), achieved impressive performance. However, they are inherently constrained in the fixed target classes, lacking the ability to generalize to more. This fundamental limitation spurred the development of Class-Agnostic Counting (CAC) (Ranjan et al. 2021), which enables counting objects of arbitrary classes at test time prompted by visual exemplars (Chen et al. 2025) or a class name (Lin and Chan 2024; Qian et al. 2025). Despite this significant leap in flexibility for counting, CAC models still operate at the class level, remaining oblivious to fine-grained attributes that distinguish different subclasses within a class. To bridge this gap, Referring Expression Counting (REC) (Dai, Liu, and Cheung 2024) was recently introduced to tackle the more challenging problem of subclass counting. REC enables the enumeration of specific object subclasses defined by textual referring expressions, marking a crucial progression towards more intelligent and practical counting systems.

Referring Expression Counting

Referring Expression Counting (REC) moves beyond class-level counting to the more challenging subclass level. The goal is to enumerate objects that match a given textual description, which specifies both a class name and the distinguishing attribute. The core challenge of REC, therefore,

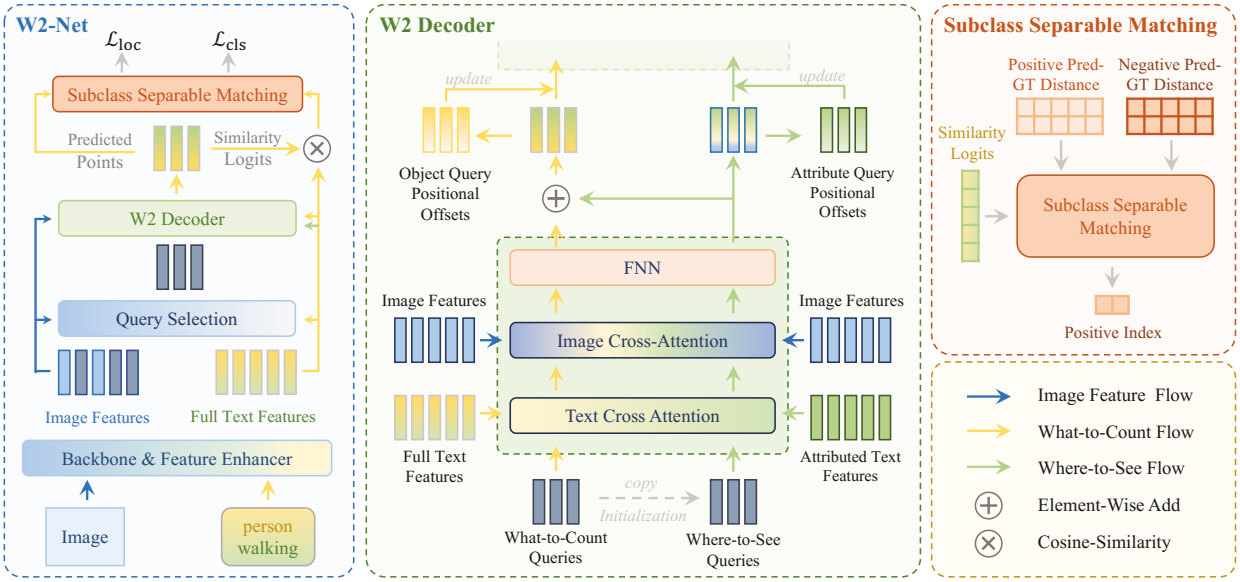

Figure 2: The framework of *W2-Net*. *W2-Net* decouples "what to count" and "where to see" in the proposed W2 Decoder, where the what-to-count (w2c) query targets at locating the object's class-representative center and the parallel dedicated where-to-see (w2s) query grounds the distinguishing attribute. Fusing their features enables precise subclass discrimination. Besides, we develop the Subclass Separable Matching (SSM) to stabilize training by introducing a repulsive force into the matching cost, effectively resolving inter-subclass ambiguity and ensuring stable supervision.

is to accurately distinguish between subclasses that share the same parent class but differ in specific attributes. The pioneering work, GroundingREC (Dai, Liu, and Cheung 2024), introduced a detection-then-count pipeline based on the open-set detector GroundingDINO (Liu et al. 2024). It enhances the detector with global-local feature fusion and contrastive learning to improve attribute differentiation. Following this, CAD-GD (Wang et al. 2025b) proposed a contextual attribute density module to guide localization. While these methods have made promising strides, they inherit and overlook a critical issue rooted in the standard annotation protocol of the REC dataset (Dai, Liu, and Cheung 2024).

The Challenge of Point Annotations in Counting

Due to their cost-effectiveness, 2D point annotations are the predominant format in object counting. In conventional class-level counting, these points are intended to mark class-representative locations (*e.g.*, persons' heads). However, they are often prone to positional errors arising from annotator subjectivity. Several methods have been proposed to mitigate this issue. For density-based counting, robust loss functions like Bayesian Loss (Ma et al. 2019) and NoiseCC (Wan, Wu, and Chan 2023) were designed to handle noisy density maps from noisy annotation points. For localization-based counting, where these density-based solutions are less applicable, SAE (Zou et al. 2024) was proposed to directly refine the point annotations by alleviating their positional inconsistency. Crucially, these methods primarily address issues related to *random positional noise*.

In REC, however, the annotation challenge transcends random noise. While REC annotations also mark class-representative locations, the supervisory signal they provide is often spatially detached from the visual regions that characterize the subclass attribute (e.g., the legs for "walking" vs the head for "person"). This misalignment forces the model to choose between focusing on the class-representative location or seeking attribute-specific regions for correct subclass distinction, creating an inherent learning conflict. Our work is the first to identify and explicitly tackle this contradiction, enabling the model to learn class features and discriminative subclass features simultaneously and effectively.

Methodology

Overview

This section details the proposed *W2-Net*, a novel framework designed to tackle a fundamental yet overlooked challenge in REC: misalignment between class-representative annotation points and attribute-defining visual regions. To address this, we introduce two synergistic modules: a novel W2 Decoder that decouples "what to count" and "where to see", and a Subclass Separable Matching (SSM) to ensure stable and accurate supervision.

Following prior methods (Dai, Liu, and Cheung 2024; Wang et al. 2025b), *W2-Net* is built upon the open-set detector, GroundingDINO (Liu et al. 2024), adapting it from box detection to point-based localization for the counting task. The overall architecture is depicted in Fig. 2. Given an input image I and a referring expression text T that specifies the class name and the attribute, we first extract image features $F_v \in \mathbb{R}^{M\times C}$ and text features $F_t \in \mathbb{R}^{N\times C}$ using a frozen backbone and a feature enhancer, where M is the number of image tokens, N is the number of text to-

kens, and C is the feature dimension. These features are then fed into our W2 Decoder. Here, dedicated *where-to-see (w2s)* queries operate in parallel with the standard *what-to-count (w2c)* queries. The w2s queries are guided to locate attribute-relevant features, which are then fused to w2c queries to enable precise subclass discrimination. Through the iterative refinement of multiple decoder layers, the enhanced w2c queries are passed to prediction heads to generate point locations and classification scores. During training, these predicted points are assigned to ground-truth labels via our Subclass Separable Matching (SSM), which introduces a repulsive force to mitigate matching ambiguity between similar subclasses. Finally, the complete model is optimized using classification and localization losses.

W2 Decoder

As previously discussed, the central challenge in REC stems from the spatial misalignment between class-representative annotation points and attribute-representative visual regions. To address this, we propose W2 Decoder, which is explicitly designed to decouple the "what to count" and "where to see". It introduces a dedicated where-to-see (w2s) query (Q_{w2s}) that operates in parallel with the standard what-to-count (w2c) query (Q_{w2c}). While the w2c query learns to converge on the annotated point ("what to count"), the w2s query is specifically guided to seek out visual evidence for the given attribute ("where to see"). This parallel, specialized processing allows the model to learn both class-level and attribute-specific features effectively.

Query Initialization. At the decoder's input (before the first decoder layer, $l = 0$), we initialize a set of K what-to-count (w2c) queries, where $K = 900$ as in GroundingDINO (Liu et al. 2024). Each w2c query consists of a content embedding and a reference point. Their content embeddings $\{q^0_{w2c,i}\}_{i=1}^K$ are learnable parameters, while their initial reference points $\{p^0_{w2c,i}\}_{i=1}^K$ are set to the spatial locations of the top-K image features with the highest cross-modal similarity to the full text feature F_t. We then initialize the where-to-see (w2s) queries by duplicating the w2c queries. Specifically, each w2s query's content embedding is a separate learnable parameter initialized from its w2c counterpart $q^0_{w2s,i} = q^0_{w2c,i}$, and its reference point is directly copied $p^0_{w2s,i} = p^0_{w2c,i}$. This co-location initialization strategy ensures that the w2s query begins its search for attribute-specific visual cues from a sensible, class-relevant starting location.

Parallel Query Refinement. Within each decoder layer l, the w2c queries and the w2s queries are refined in parallel with different focuses.

For the *w2c query*, it first attends to the full text features F_t to capture the complete semantic intent, and then to the image features F_v via deformable attention to gather visual information around its current reference point P_{w2c}:

$$Q^{txt}_{w2c} = \text{CrossAttn}(Q_{w2c}, F_t, F_t), \quad (1)$$

$$Q^{img}_{w2c} = \text{DeformAttn}(Q^{txt}_{w2c}, P_{w2c}, F_v). \quad (2)$$

The output is then processed by a Feed-Forward Network (FFN), yielding $\hat{Q}_{w2c} = \text{FFN}(Q^{img}_{w2c})$.

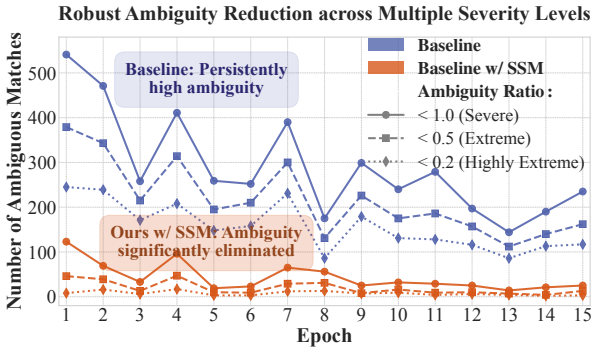

Figure 3: Effectiveness of Subclass Separable Matching (SSM) in resolving training ambiguity. A standard matching approach (blue) consistently suffers from high ambiguity due to inter-subclass similarity. In contrast, our SSM (orange), which incorporates a repulsive force, alleviates such ambiguity from the beginning of training, ensuring a stable and accurate supervision signal and improved performance.

For the *w2s query*, the process is similar but critically modified to isolate attribute-specific text features. It attends to a masked version of the text features, where class-related tokens are suppressed. Let M_{w2s} be a binary mask that is one for attribute tokens and zero otherwise. This mask is straightforwardly derived from the provided annotations, which explicitly specify the class word for each referring expression, thus incurring no additional overhead. The query is refined as:

$$Q^{txt}_{w2s} = \text{CrossAttn}(Q_{w2s}, F_t \odot M_{w2s}, F_t \odot M_{w2s}), \quad (3)$$

$$Q^{img}_{w2s} = \text{DeformAttn}(Q^{txt}_{w2s}, P_{w2s}, F_v). \quad (4)$$

This forces the w2s query to focus exclusively on attribute-related words (*e.g.*, "walking"), guiding it to search for corresponding visual evidence specifically related to the attribute. This also yields an intermediate representation $\hat{Q}_{w2s} = \text{FFN}(Q^{img}_{w2s})$.

Fusion and Iterative Update. After parallel refinement, the attribute-centric information from the w2s query is injected into the w2c query stream to enhance its discriminative ability. This fusion is achieved via a simple element-wise addition. The fused representation is then passed through an FFN to produce the updated w2c query for the next layer $(l + 1)$:

$$Q^{l+1}_{w2c} = \text{FFN}(\hat{Q}^l_{w2c} + \hat{Q}^l_{w2s}). \quad (5)$$

Crucially, the w2s query is updated independently, preserving its specialized role:

$$Q^{l+1}_{w2s} = \text{FFN}(\hat{Q}^l_{w2s}). \quad (6)$$

This independent update ensures that the w2s query remains a dedicated scout for attribute regions in the next layer, preventing its focus from being overly diluted by the w2c query's primary goal of localizing the object.

Finally, the reference points are updated. Two separate localization heads predict positional offsets from the refined queries, which are then added to the current reference points

Method	Venue	Backbone	Validation Set				
			MAE ↓	RMSE ↓	Prec ↑	Rec ↑	F1 ↑
ZSC (Xu et al. 2023)	CVPR23	Swin-T	12.96	26.74	-	-	-
CounTX (Amini-Naieni et al. 2023)	BMVC23	ViT-B	11.88	27.04	-	-	-
GroundingDINO (Liu et al. 2024)	ECCV24	Swin-T	9.03	21.98	0.56	<u>0.76</u>	0.65
GroundingREC (Dai, Liu, and Cheung 2024)	CVPR24	Swin-T	6.80	18.13	0.65	0.71	0.68
CAD-GD (Wang et al. 2025b)	CVPR25	Swin-T	5.43	15.01	0.68	0.72	0.70
CAD-GD† (Wang et al. 2025b)	CVPR25	Swin-T	<u>4.58</u>	<u>13.24</u>	<u>0.68</u>	0.71	<u>0.70</u>
W2-Net (Ours)	-	Swin-T	**3.55**	**10.39**	**0.76**	**0.78**	**0.77**
GroundingREC (Dai, Liu, and Cheung 2024)	CVPR24	Swin-B	5.66	15.24	0.66	<u>0.77</u>	0.71
CAD-GD (Wang et al. 2025b)	CVPR25	Swin-B	4.83	13.52	0.74	0.76	<u>0.75</u>
CAD-GD† (Wang et al. 2025b)	CVPR25	Swin-B	<u>4.23</u>	<u>13.14</u>	<u>0.76</u>	0.70	0.73
W2-Net (Ours)	-	Swin-B	**3.47**	**9.55**	**0.79**	**0.79**	**0.79**

Table 1: Comparison with state-of-the-art approaches on the REC-8K validation set (Dai, Liu, and Cheung 2024). The best and second best are boldfaced and <u>underlined</u>, respectively. † indicates using a density-based query selection strategy.

Method	Venue	Backbone	Test Set				
			MAE ↓	RMSE ↓	Prec ↑	Rec ↑	F1↑
ZSC (Xu et al. 2023)	CVPR23	Swin-T	13.00	29.07	-	-	-
CounTX (Amini-Naieni et al. 2023)	BMVC23	ViT-B	11.84	25.62	-	-	-
GroundingDINO (Liu et al. 2024)	ECCV24	Swin-T	8.88	21.95	0.59	<u>0.76</u>	0.66
GroundingREC (Dai, Liu, and Cheung 2024)	CVPR24	Swin-T	6.50	19.79	0.67	0.72	0.69
CAD-GD (Wang et al. 2025b)	CVPR25	Swin-T	5.29	17.08	0.71	0.73	<u>0.72</u>
CAD-GD† (Wang et al. 2025b)	CVPR25	Swin-T	<u>4.59</u>	<u>14.68</u>	<u>0.72</u>	0.70	0.71
W2-Net (Ours)	-	Swin-T	**3.59**	**9.59**	**0.77**	**0.80**	**0.79**
GroundingREC (Dai, Liu, and Cheung 2024)	CVPR24	Swin-B	5.42	18.47	0.71	0.69	0.70
CAD-GD (Wang et al. 2025b)	CVPR25	Swin-B	4.94	14.65	0.75	<u>0.77</u>	<u>0.76</u>
CAD-GD† (Wang et al. 2025b)	CVPR25	Swin-B	<u>4.34</u>	<u>12.93</u>	<u>0.77</u>	0.71	0.74
W2-Net (Ours)	-	Swin-B	**3.56**	**9.15**	**0.79**	**0.81**	**0.80**

Table 2: Comparison with state-of-the-art approaches on the REC-8K test set. (Dai, Liu, and Cheung 2024) The best and second best are boldfaced and <u>underlined</u>, respectively. † indicates using a density-based query selection strategy.

to obtain P_{w2c}^{l+1} and P_{w2s}^{l+1} for the next iteration. This iterative process allows w2c queries to converge on the annotated centers while w2s queries simultaneously navigate towards attribute-defining regions.

Subclass Separable Matching

A standard one-to-one Hungarian matching (Kuhn 1955) based on classification and localization costs is suboptimal for REC. The high visual similarity between different subclasses (*e.g.*, "person walking" vs. "person standing") often leads to matching ambiguity, where a query corresponding to an object of another subclass might be incorrectly matched to a ground truth, especially in early training stages (see Fig. 3). This results in unstable and noisy supervision.

To mitigate this, we propose Subclass Separable Matching (SSM), which introduces a repulsive force into the matching cost. This force explicitly penalizes a query assignment if the query is close to ground-truth points of other, non-target subclasses. Let $Y_{pos} = \{p_j\}_{j=1}^{N_{pos}}$ be the set of ground-truth points for the target subclass, and $Y_{neg} = \{p_k\}_{k=1}^{N_{neg}}$ be the ground-truth points for all other subclasses in the image. The matching cost $\mathcal{C}_{match}(i,j)$ between prediction i and a ground-truth point $j \in Y_{pos}$ is:

$$\begin{aligned}\mathcal{C}_{match}(i,j) = \lambda_{cls}(1-\hat{s}_i) + \lambda_{L1}\|\hat{p}_i - p_j\|_1 \\ + \lambda_{rep}\mathcal{C}_{rep}(\hat{p}_i, p_j),\end{aligned} \tag{7}$$

where $\hat{s}_i$ and $\hat{p}_i$ are the predicted score and location, and λs are balancing weights (set to $\lambda_{cls} = 5$ and $\lambda_{L1} = 1$ as in prior works GroundingREC (Dai, Liu, and Cheung 2024) and CAD-GD (Wang et al. 2025b), and a hyperparameter λ_{rep}, respectively). The key component is our proposed repulsive cost $\mathcal{C}_{rep}$, which penalizes a query assignment if it is closer to a non-target subclass. To achieve this, we first define an ambiguity ratio, $\mathcal{R}(\hat{p}_i, p_j)$, which measures the relative proximity of a prediction $\hat{p}_i$ to the nearest non-target object versus its potential ground-truth match p_j:

$$\mathcal{R}(\hat{p}_i, p_j) = \frac{\min_{p_k \in Y_{neg}} \|\hat{p}_i - p_k\|_2}{\|\hat{p}_i - p_j\|_2} \,. \tag{8}$$

The repulsive cost $\mathcal{C}_{rep}$ is then calculated based on this ratio:

$$\mathcal{C}_{rep}(\hat{p}_i, p_j) = \exp(-\mathcal{R}(\hat{p}_i, p_j)) \,. \tag{9}$$

The cost becomes large if a prediction $\hat{p}_i$ is closer to a negative object of another subclass (a distractor) than to its potential positive match p_j, thus discouraging ambiguous assignments. By incorporating this repulsive force, our matching

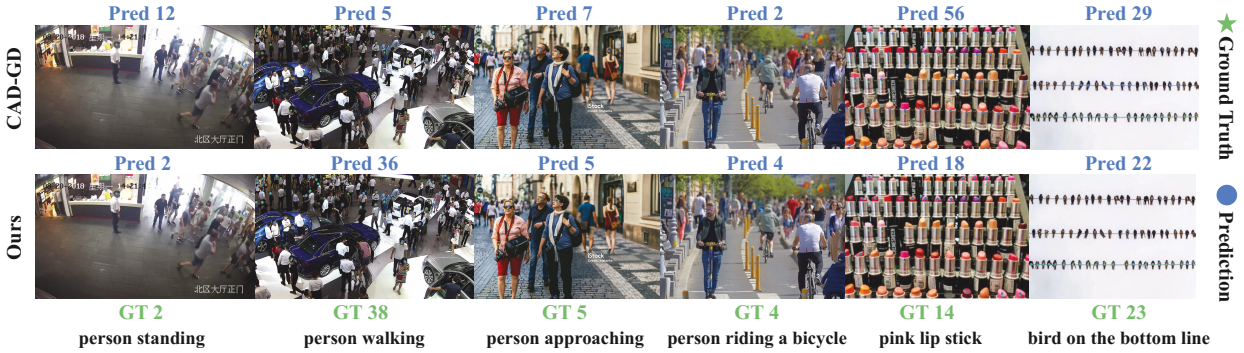

Figure 4: Some qualitative results on the REC-8K dataset of CAD-GD (Wang et al. 2025b) and our W2-Net.

enforces greater inter-subclass separability during the crucial label assignment phase, leading to more stable training and more discriminative final representations. As illustrated in Fig. 3, our SSM effectively eliminates matching ambiguities in training, ensuring a more stable learning process.

Training Objective

Following the one-to-one label assignment by our Subclass Separable Matching, the network is optimized using a loss function that mirrors GroundingREC and CAD-GD. The total loss $\mathcal{L}$ is a weighted sum of a classification loss $\mathcal{L}_{cls}$ and a localization loss $\mathcal{L}_{loc}$:

$$\mathcal{L} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{L1}\mathcal{L}_{loc}, \tag{10}$$

where the weights λ_{cls} and λ_{loc} are consistent with those in the matching cost, set to $\lambda_{cls} = 5$ and $\lambda_{L1} = 1$ as in prior works. Specifically, we employ the focal loss for $\mathcal{L}_{cls}$ to handle the class imbalance between positive and negative queries, and the L1 loss for $\mathcal{L}_{loc}$ to point localization.

Experiments

Experimental Setup

Dataset and Metrics. REC-8K (Dai, Liu, and Cheung 2024), the first REC benchmark, with 8,011 images and 17,122 image-RE pairs. Following prior work (Dai, Liu, and Cheung 2024; Wang et al. 2025b), we use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for counting, and Precision, Recall, and F1-score for localization. Localization metric as it verifies precise subclass discrimination, preventing deceptively low counting errors from error cancellation.

Implementation Details. We train W2-Net for 15 epochs on REC-8K using the AdamW optimizer with a learning rate of 10^{-5}. Each training batch uses a single image with its corresponding referring expressions.

Comparison with State-of-the-Art

We compare our *W2-Net* with existing methods on the REC-8K benchmark in Tab.1 and Tab.2. *W2-Net* establishes **new state-of-the-art (SOTA)**, significantly outperforming all prior works on both counting and localization metrics.

With a Swin-T backbone (Liu et al. 2021), *W2-Net* reduces the SOTA MAE from 4.58 to 3.55 on the validation set and from 4.59 to 3.59 on the test set, corresponding to a relative error reduction of 22.5% and 18.0%, respectively. Concurrently, it boosts the localization F1-score by a significant 7%. The performances are more pronounced with a Swin-B backbone, where our method achieves a new record with an MAE of 3.56 and an F1-score of 0.80 on the test set, substantially widening the gap with previous methods.

The substantial gains confirm the success of our approach in tackling the identified annotation challenge. By enabling the model to jointly reason about class-representative locations ("what to count") and attribute-specific visual regions ("where to see"), *W2-Net* learns a highly discriminative representation with more stable supervision provided by our Subclass Separable Matching (SSM). This directly translates into reduced subclass confusion and superior localization accuracy, leading to more reliable counting performance and setting a new record for the REC task. Fig. 4 provides qualitative examples, visually demonstrating that *W2-Net* successfully distinguishes challenging subclasses.

W2 Decoder	SSM	MAE	RMSE	Prec	Rec	F1
		7.21	18.69	0.63	0.69	0.66
	✓	5.93	16.71	0.66	0.71	0.68
✓		3.85	11.55	0.75	0.77	0.76
✓	✓	**3.55**	**10.39**	**0.76**	**0.78**	**0.77**

Table 3: Ablation study on the proposed W2 Decoder and Subclass Separable Matching (SSM).

λ_{rep}	0	0.1	0.2	0.5	1.0
MAE ↓	3.85	3.67	**3.55**	3.64	3.92
F1 ↑	0.76	0.76	**0.77**	0.77	0.74

Table 4: Analysis on repulsive weight λ_{rep} in SSM.

Ablation Study

All ablation studies are conducted on the REC-8K validation set using the Swin-T backbone.

Proposed Components. We first evaluate the impact of our two main contributions independently. As shown in Tab. 3, starting from a baseline MAE of 7.21, integrating our Subclass Separable Matching (SSM) alone reduces the error to

Method	Venue	Prompt	Val Set		Test Set	
			MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
LOCA (Djukic et al. 2023)	ICCV23	Visual	10.24	32.56	10.79	56.97
CACViT (Wang et al. 2024)	AAAI24	Visual	10.63	37.95	9.13	48.96
GeCo (Pelhan et al. 2024a)	NIPS24	Visual	9.52	43.00	7.91	54.28
DAVE (Pelhan et al. 2024b)	CVPR24	Visual	8.91	28.08	8.66	32.36
CountGD (Amini-Naieni, Han, and Zisserman 2024)	NIPS24	Visual	**7.10**	**26.08**	**5.74**	**24.09**
Patch-Selection (Xu et al. 2023)	CVPR23	Text	26.93	88.63	22.09	115.17
CLIP-Count (Jiang, Liu, and Chen 2023)	ACMMM23	Text	18.79	61.18	17.78	106.62
VLCounter (Kang et al. 2024)	AAAI24	Text	18.06	65.13	17.05	106.16
CounTX (Amini-Naieni et al. 2023)	BMVC23	Text	17.10	65.61	15.88	106.29
DAVE (Pelhan et al. 2024b)	CVPR24	Text	15.48	52.57	14.90	103.42
GroundingREC (Dai, Liu, and Cheung 2024)	CVPR24	Text	10.06	58.62	10.12	107.19
CountGD (Amini-Naieni, Han, and Zisserman 2024)	NIPS24	Text	12.14	47.51	12.98	98.35
CAD-GD (Wang et al. 2025b)	CVPR25	Text	9.30	40.96	10.35	86.88
W2-Net (Ours)	-	Text	**8.73**	**37.32**	**9.53**	**83.46**

Table 5: Comparison with some state-of-the-art approaches on FSC-147 (Ranjan et al. 2021). The upper and lower parts present the results using visual exemplars and class text as prompts separately.

Method	Prompt	Test Set	
		MAE ↓	RMSE ↓
CLIP-Count	Text	11.96	16.61
CounTX*	Text	8.13	10.87
VLCounter	Text	6.46	8.68
CACViT*	Visual	4.91	6.49
CountGD	Text	3.83	5.41
CountGD	Both	3.68	5.17
CAD-GD	Text	3.29	4.56
W2-Net (Ours)	Text	**2.89**	**4.13**

Table 6: Comparison with some state-of-the-art approaches on the CARPK dataset. * means finetuned.

5.93 by stabilizing the training signal. Independently, the W2 Decoder yields a more substantial improvement (MAE 3.85, F1 0.76), confirming that explicitly modeling "where to see" is crucial for fine-grained localization. Combining both components in our full *W2-Net* model yields the best performance, demonstrating their synergistic effect: the W2 Decoder provides superior features, while SSM ensures they are learned through precise label assignment.

Analysis of Repulsive Weight λ_{rep} in Eq. 7. We analyze the effect of the repulsive force hyperparameter λ_{rep} in SSM. As presented in Tab. 4, performance improves as λ_{rep} increases from 0 (disabling the repulsive force), peaking at $\lambda_{rep} = 0.2$. A further increase slightly degrades performance, suggesting that an overly strong repulsive force can incorrectly overwhelm other matching cost terms, such as the classification. We thus set $\lambda_{rep} = 0.2$.

Results on Zero-shot Counting

FSC-147 (Ranjan et al. 2021). To assess the generalizability of our framework, we evaluate *W2-Net* on the FSC-147 benchmark for zero-shot class-agnostic counting. As shown in Tab. 5, our method achieves strong performance, outperforming previous text-prompted approaches with an MAE of 8.73 and 9.53 on the validation and test sets, respectively. It is crucial to note that FSC-147 provides only class-level labels (*e.g.*, "car"), lacking the fine-grained attributes needed to explicitly guide our w2s query. The strong performance, therefore, highlights the inherent robustness of our dual-query design. Even without explicit attribute supervision, the w2s query acts as a dynamic feature scout, automatically exploring supplementary visual cues that enrich the w2c query's representation. This capability leads to more robust class-level identification.

CARPK (Hsieh, Lin, and Hsu 2017). To further test the cross-dataset generalization, we directly evaluate our FSC-147-trained model on the CARPK car counting dataset without any fine-tuning. The results, presented in Tab.6, demonstrate that *W2-Net* surpasses existing methods, including those that leverage visual exemplar prompts, demonstrating its excellent transferability.

Conclusion

In this paper, we identify and address a critical, yet overlooked, challenge in Referring Expression Counting (REC): the misalignment between class-representative annotations and attribute-defining visual regions. We introduce *W2-Net*, a novel framework that resolves this issue by decoupling "what to count" and "where to see" via its W2 decoder. This approach, synergized with our Subclass Separable Matching, significantly outperforms previous REC methods in counting and localization. Specifically, *W2-Net* achieves a remarkable reduction in relative counting error by 22.5% on the validation set and 18.0% on the test set, while also significantly improving the localization accuracy. Beyond these impressive performance gains, we hope this work can shed light on a fundamental annotation challenge in REC.

Acknowledgments

This work was supported in part by the NSFC 62222112, 62176186, and the Innovative Research Group Project of Hubei Province (2024AFA017).

References

Amini-Naieni, N.; Amini-Naieni, K.; Han, T.; and Zisserman, A. 2023. Open-world Text-specified Object Counting. In *Proc. of British Machine Vision Conference*.

Amini-Naieni, N.; Han, T.; and Zisserman, A. 2024. Countgd: Multi-modal open-world counting. In *Proc. of Advances in Neural Information Processing Systems*, volume 37, 48810–48837.

Chen, X.; Huo, S.; Jiang, B.; Hu, H.; and Chen, X. 2025. Single Domain Generalization for Few-Shot Counting via Universal Representation Matching. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 4639–4649.

Chen, Y.; Liang, D.; Bai, X.; Xu, Y.; and Yang, X. 2021. Cell localization and counting using direction field map. *IEEE Journal of Biomedical and Health Informatics*, 26(1): 359–368.

Dai, S.; Liu, J.; and Cheung, N.-M. 2024. Referring expression counting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 16985–16995.

Djukic, N.; Lukezic, A.; Zavrtanik, V.; and Kristan, M. 2023. A low-shot object counting network with iterative prototype adaptation. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 18872–18881.

Dumery, C.; Etté, N.; Fan, A.; Li, R.; Xu, J.; Le, H.; and Fua, P. 2025. Counting Stacked Objects. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*.

Guo, H.; Gao, J.; and Yuan, Y. 2025. Enhancing Low-Rank Adaptation with Recoverability-Based Reinforcement Pruning for Object Counting. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 39, 3238–3246.

Han, T.; Bai, L.; Liu, L.; and Ouyang, W. 2023. STEERER: Resolving Scale Variations for Counting and Localization via Selective Inheritance Learning. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 21848–21859.

He, Y.; Dai, Z.; Trigoni, N.; Chen, L.; and Markham, A. 2024. SoundCount: sound counting from raw audio with dyadic decomposition neural network. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 38, 12421–12429.

Hsieh, M.-R.; Lin, Y.-L.; and Hsu, W. H. 2017. Drone-based object counting by spatially regularized regional proposal network. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 4145–4153.

Huang, Y.; Nguyen, D. D.; Nguyen, L.; Pham, C.; and Hoai, M. 2024. Count what you want: exemplar identification and few-shot counting of human actions in the wild. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 38, 10057–10065.

Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Maadeed, S.; Rajpoot, N.; and Shah, M. 2018. Composition loss for counting, density map estimation and localization in dense crowds. In *Proc. of European Conf. on Computer Vision*, 532–546.

Jiang, R.; Liu, L.; and Chen, C. 2023. Clip-count: Towards text-guided zero-shot object counting. In *Proc. of the ACM International Conference on Multimedia*, 4535–4545.

Kang, S.; Moon, W.; Kim, E.; and Heo, J.-P. 2024. Vlcounter: Text-aware visual representation for zero-shot object counting. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 38, 2714–2722.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2): 83–97.

Li, T.; Chang, H.; Wang, M.; Ni, B.; Hong, R.; and Yan, S. 2014. Crowded scene analysis: A survey. *IEEE Trans. on Circuits and Systems for Video Technology*, 25(3): 367–386.

Liang, D.; Xu, W.; and Bai, X. 2022. An end-to-end transformer model for crowd localization. In *Proc. of European Conf. on Computer Vision*, 38–54.

Lin, H.; Ma, Z.; Hong, X.; Shangguan, Q.; and Meng, D. 2024. Gramformer: Learning Crowd Counting via Graph-Modulated Transformer. In *Proc. of the AAAI Conf. on Artificial Intelligence*.

Lin, W.; and Chan, A. B. 2024. A fixed-point approach to unified prompt-based counting. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 38, 3468–3476.

Lin, W.; Zhao, C.; and Chan, A. B. 2025. Point-to-Region Loss for Semi-Supervised Point-Based Crowd Counting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 29363–29373.

Liu, C.; Lu, H.; Cao, Z.; and Liu, T. 2023. Point-Query Quadtree for Crowd Counting, Localization, and More. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 1676–1685.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proc. of European Conf. on Computer Vision*, 38–55. Springer.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 10012–10022.

Ma, Z.; Wei, X.; Hong, X.; and Gong, Y. 2019. Bayesian loss for crowd count estimation with point supervision. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 6142–6151.

Mondal, A.; Nag, S.; Zhu, X.; and Dutta, A. 2025. Omnicount: Multi-label object counting with semantic-geometric priors. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 39, 19537–19545.

Pelhan, J.; Lukezic, A.; Zavrtanik, V.; and Kristan, M. 2024a. A Novel Unified Architecture for Low-Shot Counting by Detection and Segmentation. In *Proc. of Advances in Neural Information Processing Systems*, volume 37, 66260–66282.

Pelhan, J.; Zavrtanik, V.; Kristan, M.; et al. 2024b. DAVE-A Detect-and-Verify Paradigm for Low-Shot Counting. In

Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 23293–23302.

Perez, G.; Maji, S.; and Sheldon, D. 2024. DISCount: counting in large image collections with detector-based importance sampling. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 38, 22294–22302.

Pothiraj, A.; Stengel-Eskin, E.; Cho, J.; and Bansal, M. 2025. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*.

Qian, Y.; Guo, Z.; Deng, B.; Lei, C. T.; Zhao, S.; Lau, C. P.; Hong, X.; and Pound, M. P. 2025. T2icount: Enhancing cross-modal understanding for zero-shot counting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 25336–25345.

Ranjan, V.; Sharma, U.; Nguyen, T.; and Hoai, M. 2021. Learning to count everything. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 3394–3403.

Shi, M.; Zhang, X.; Yue, Z.; Luo, Y.; Zhao, C.; and Li, L. 2025. Text-promptable Object Counting via Quantity Awareness Enhancement. arXiv:2507.06679.

Sindagi, V. A.; Yasarla, R.; and Patel, V. M. 2020. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 44(5): 2594–2609.

Song, Q.; Wang, C.; Jiang, Z.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Wu, Y. 2021. Rethinking counting and localization in crowds: A purely point-based framework. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 3365–3374.

Sun, G.; An, Z.; Liu, Y.; Liu, C.; Sakaridis, C.; Fan, D.-P.; and Van Gool, L. 2023. Indiscernible Object Counting in Underwater Scenes. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 13791–13801.

Triaridis, K.; Kaliosis, P.; Nguyen, E.-R.; Xu, J.; Le, H.; and Samaras, D. 2025. Improving Contrastive Learning for Referring Expression Counting. arXiv:2505.22850.

Wan, J.; Wu, Q.; and Chan, A. B. 2023. Modeling Noisy Annotations for Point-Wise Supervision. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 45(12): 15065–15080.

Wang, M.; Zhou, J.; Dai, Y.; Buys, E.; and Gong, M. 2025a. Enhancing zero-shot counting via language-guided exemplar learning. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*.

Wang, Z.; Pan, Z.; Peng, Z.; Cheng, J.; Xiao, L.; Jiang, W.; and Cao, Z. 2025b. Exploring Contextual Attribute Density in Referring Expression Counting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*.

Wang, Z.; Xiao, L.; Cao, Z.; and Lu, H. 2024. Vision transformer off-the-shelf: a surprising baseline for few-shot class-agnostic counting. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 38, 5832–5840.

Xu, J.; Le, H.; Nguyen, V.; Ranjan, V.; and Samaras, D. 2023. Zero-shot object counting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 15548–15557.

Yang, C.; Geng, T.; Peng, J.; and Xu, C. 2025a. PBECount: Prompt-Before-Extract Paradigm for Class-Agnostic Counting. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 39, 9139–9147.

Yang, M.; Li, Z.; Zhang, J.; Qi, L.; and Shi, Y. 2025b. Taste More, Taste Better: Diverse Data and Strong Model Boost Semi-Supervised Crowd Counting. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 24440–24451.

Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; and Ma, Y. 2016. Single-image crowd counting via multi-column convolutional neural network. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 589–597.

Zou, Y.; Xiao, X.; Zhou, P.; Sun, Z.; Du, B.; and Xu, Y. 2024. Shifted autoencoders for point annotation restoration in object counting. In *Proc. of European Conf. on Computer Vision*, 113–130. Springer.