Improving Contrastive Learning for Referring Expression Counting

Kostas Triaridis[1*] Panagiotis Kaliosis[1*] E-Ro Nguyen[1]
Jingyi Xu[1] Hieu Le[2] Dimitris Samaras[1]

[1]Stony Brook University [2]EPFL

Abstract

Object counting has progressed from class-specific models, which count only known categories, to class-agnostic models that generalize to unseen categories. The next challenge is Referring Expression Counting (REC), where the goal is to count objects based on fine-grained attributes and contextual differences. Existing methods struggle with distinguishing visually similar objects that belong to the same category but correspond to different referring expressions. To address this, we propose C-REX, a novel contrastive learning framework, based on supervised contrastive learning, designed to enhance discriminative representation learning. Unlike prior works, C-REX operates entirely within the image space, avoiding the misalignment issues of image-text contrastive learning, thus providing a more stable contrastive signal. It also guarantees a significantly larger pool of negative samples, leading to improved robustness in the learned representations. Moreover, we showcase that our framework is versatile and generic enough to be applied to other similar tasks like class-agnostic counting. To support our approach, we analyze the key components of sota detection-based models and identify that detecting object centroids instead of bounding boxes is the key common factor behind their success in counting tasks. We use this insight to design a simple yet effective detection-based baseline to build upon. Our experiments show that C-REX achieves state-of-the-art results in REC, outperforming previous methods by more than 22% in MAE and more than 10% in RMSE, while also demonstrating strong performance in class-agnostic counting. Code is available at this url.

This paper is currently under review at the International Conference on Computer Vision (ICCV 2025).

A pre-print is available at this url.

Keywords: *Vision and Learning.*

*Equal contribution. Correspondence to kostas@cs.stonybrook.edu

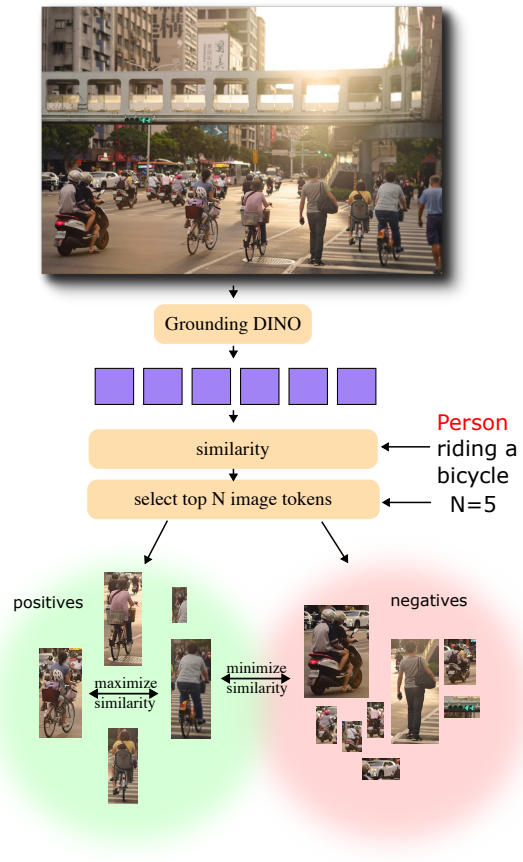

Figure 1. Our proposed method C-REX. C-REX aligns embeddings of objects sharing the same class and referring expression while separating those with different expressions or classes.

1. Introduction

Object counting methods aim to predict the number of instances of objects in some specified category in an image. More recent methods have shifted focus from class-specific counting, where the model is able to count only known categories, to class-agnostic counting, where the model can

handle categories unseen during training. The next logical step for designing more general and robust counting models is enabling them to differentiate between instances of categories with specific attributes and context within an image and count them separately. To this end Dai et al. [12] introduced the task of Referring Expression Counting (REC), which aims to count instances of objects with fine-grained contextual differences *e.g.* "person walking/standing" or "box containing grapes/mango slices".

A fundamental challenge in Referring Expression Counting (REC) is distinguishing between visually similar objects that belong to the same category but correspond to different referring expressions. Traditional counting methods struggle in this setting, as they often lack the ability to disambiguate objects based on nuanced differences in attributes. Ensuring that the final count reflects only the correct subset of objects requires models that have learned discriminative features that go beyond basic object recognition and can instead accurately represent the fine-grained differences within the same class. To achieve this, we introduce **C-REX** (Contrastive Learning for Referring Expressions) a novel Contrastive Learning (CL) approach, based on supervised contrastive learning [25]. C-REX brings together image embeddings of objects from the same class that also correspond to the same referring expression while pushing away those tied to different expressions or different classes. Concretely, we designate as "positive" samples the N image embeddings most similar to the referring expression and treat the rest as negatives, where N is the ground truth object count within an image for a referring expression. We also confirm that using the ground truth object count as the number of positive selections is a justified choice, both theoretically and through experimental results (Section 4.5).

A key strength of our contrastive learning framework is that it is **general** and **versatile**, as it can be applied to any task that requires learning discriminative visual features for objects with subtle, context-dependent differences like those described by referring expressions. In this spirit, we make a simple change to adapt our approach for class-agnostic counting, for which we select as positives the samples that are more similar to the given *class*, since there are no referring expressions. With this adaptation, our model achieves competitive performance in the class-agnostic counting task, outperforming previous text-based methods, as shown in Section 4.

We highlight that C-REX improves upon Dai et al. [12] by addressing key limitations, in the context of contrastive learning. By operating entirely within the image space rather than contrasting image and text embeddings, it ensures a more stable contrastive signal. Additionally, it provides a significantly larger pool of negative samples, as the number of candidate image tokens for detection consistently exceeds the number of referring expressions for a category in a given image.

We begin by developing a simple detection-based baseline to implement our novel contrastive paradigm, as it offers localization and explainability which are key advantages over density-based models, despite the latter historically achieving better performance. These factors are especially crucial for Referring Expression Counting (REC), where visually similar instances must be distinguished [15, 38]. To understand the recent success of detection-based methods, we analyze key components of state-of-the-art approaches like CountGD [3] and GroundingREC [12], which have outperformed density-based models. We identify that the common key factor behind their success is the re-purposing of robust open-set detectors like Grounding DINO [29] from bounding box to object centroid predictors, as this allows them to more robustly identify objects in dense and cluttered scenes. Our experiments in Section 4.5 validate this insight. Building on this observation, we design a simple yet effective detection-based baseline that achieves strong performance in both class-agnostic counting and REC.

To summarize, our contributions are the following:

- We identify the conversion of modern open-set detection models to centroid predictors as the key component that has allowed detection-based methods to reach sota performance in counting tasks. Based on this insight, we design a new baseline to serve as the foundation of our models, that also achieves competitive performance.
- Building on this baseline, we propose C-REX, a novel contrastive learning method for REC based on supervised contrastive learning, that achieves state-of-the-art performance, outperforming previous works by over 1.4 points in MAE and over 2 points in RMSE.
- We show that C-REX is general and versatile; it can be adapted for other visual tasks that require discriminating between visually similar but contextually distinct objects. We adapt it for class-agnostic counting, and demonstrate that its performance is on par with the best text-based counting models.

2. Related work

Early works in object counting focused mostly on class-specific counting [1, 49, 61], addressing challenges in diverse domains such as crowd counting [13, 30, 41, 48] or traffic analysis [5]. More recent work has moved to class-agnostic counting(CAC)[42, 58, 62] where the aim is to generalize counting across various object categories, typically not encountered during training [19, 32, 46, 51, 54]. To enable this, many methods used textual description to describe the classes [2, 8, 24], while others use a small number of annotated examples as visual reference [21, 22, 34, 56] to achieve better results. More recently Dai et al. [12] introduced Referring Expression Counting (REC), where the

goal is to count only the subsets of instances of a class in an image that match a given referring expression. Each referring expression consists of a **class**, specifying the object category, and an **attribute**, describing its fine-grained characteristics. REC is closely related to Referring Expression Segmentation [33, 50, 53] and Referring Expression Comprehension [39, 45], which focus on segmentation or localization rather than counting.

Early approaches on object counting mostly relied on density map regression [4, 26, 27] or bounding box detection [11, 16] in order to predict the object count. Density-based methods estimate the final count by summing over a density map, being traditionally more accurate in cluttered and densely populated scenes where detection-based methods struggle [3]. However, a critical limitation of density-based methods is the lack of object-level detail which limits their applicability in cases that require localization and explainability [15, 38]. This is especially important in REC, where images contain multiple visually similar instances of the same category, each associated with a different referring expression. In such cases, ensuring that the final count accurately reflects only the correct subset of objects is crucial for reliable and precise counting. Recent approaches, such as CountGD [3] and GeCo [37], leverage state-of-the-art open-set object detectors [29, 60], and reach performance on-par with the best density-based methods [36].

For REC it is crucial to be able to disambiguate between visually similar objects with subtle contextual differences. For this reason, Dai et al. [12] introduce a CL module that contrasts image and referring expression embeddings. Contrastive learning in a language-image setting generally requires more data and larger-scale training to achieve robust representations [10, 40], whereas contrastive learning in the image space is more efficient [35, 60]. This is evident when comparing the performance of text-image models like CLIP [40] to image-specific models like DINOv2 [35] on smaller datasets, where DINOv2 achieves comparable performance to CLIP even when trained on significantly less data. Additionally, their approach cannot utilize a large batch size with numerous negative samples, which is crucial for improving the robustness of the learned representations [9, 17, 18, 25, 47]. This motivates our approach, which operates in image space only and can leverages a large amount of negative samples and outperforms [12] in REC.

3. Method

3.1. New Baseline for Detection-based Counting

Historically, density-based methods have showcased better performance than detection-based methods, as they were more accurate when counting large numbers of instances, and in scenes with cluttered and dense objects. However, recent detection-based methods have started becoming prominent again as works like CountGD [3] and GroundingREC [12] achieved state-of-the-art results in different counting benchmarks.

One advantage of detection-based counting is that it inherently provides **explainability** and **localization**, as each count corresponds to a specific detection, making it straightforward to identify which instances are being counted. This is particularly crucial in the context of Referring Expression Counting (REC), where images usually contain multiple visually similar instances of a single category that correspond to different referring expressions. In this scenario it is essential to verify that the final count reflects only the correct subset of these objects.

To this end, we identify two key components of the recent detection-based works [3, 12] that enable their state-of-the-art performance: the use of new and robust open-set detectors like Grounding DINO [29] and their conversion to object-centroid predictors instead of bounding box detectors. We use this observation to design a simple new detection-based baseline for object-counting that demonstrates competitive performance in both class-agnostic and referring-expression counting. Specifically, we finetune the original Grounding DINO architecture with two losses: an L1 loss for point center regression and a cross-entropy classification loss, essentially replacing bounding box prediction with point center prediction for Grounding DINO. In Section 4 we show that even this simple baseline achieves performance comparable to the state of the art, especially in the task of referring expression counting (Table 1). We refer to this baseline as GDino improved.

3.2. Contrastive Learning for REC

For referring expression counting it is important to be able to robustly differentiate objects that belong in the same class but have different attributes (referring expressions). To this end we propose a novel contrastive learning module, that learns more robust discriminative features using only the referring expression and its corresponding ground truth object count N as supervision.

Our goal is to perform contrastive learning within the same image space, as text-image contrastive learning tends to be less stable. In the context of REC, our approach must accommodate multiple positive samples, since an image typically contains multiple instances corresponding to the same class-referring expression pair. To ensure effective representation learning, we aim to bring the embeddings of these instances closer together while maintaining clear separation from other objects. For this reason our proposed loss is based on the supervised contrastive loss [25] that extends the standard contrastive loss to be able to handle multiple positive samples for each anchor. Given a set of samples I, and a set of positives $p \in P(i)$ for each sample i the supervised contrastive loss is formulated as follows:

Method	Backbone	FT	REC	Val set			Test set		
				MAE↓	RMSE↓	F1↑	MAE↓	RMSE↓	F1↑
Mean	-	✗	✗	14.28	27.75	-	13.75	25.91	-
ZSC [55]	ResNet-50	✓	✗	14.84	31.30	-	14.93	29.72	-
ZSC [55]	Swin-T	✓	✗	12.96	26.74	-	13.00	29.07	-
TFOC [58]	ViT-B	✗	✗	16.08	31.61	0.12	17.27	32.68	0.11
CounTX [2]	ViT-B-16	✓	✗	11.88	27.04	-	11.84	25.62	-
CountGD [3]	Swin-B	✓	✗	9.51	22.91	-	11.33	30.87	-
GDino [29]	Swin-T	✓	✗	9.03	21.98	0.65	8.88	21.95	0.66
GroundingREC [12]	Swin-T	✓	✓	6.80	18.13	0.68	6.50	19.79	0.69
GDino improved	Swin-T	✓	✗	5.92	17.09	0.65	5.90	19.73	0.68
C-REX (ours)	Swin-T	✓	✓	**4.86**	**13.60**	**0.68**	**5.06**	**17.53**	**0.70**

Table 1. Comparison of results on the REC-8K dataset. GDino is short for GroundingDino. Results are obtained from Dai et al. [12], except for CountGD which we retrained on the REC-8K dataset using their publicly available code. FT indicates whether the model was fine-tuned on REC-8K, while REC refers to models that were specifically designed for the task. Best results are in bold.

$$\mathcal{L}_{\text{sup}} = \sum_{i \in I} -\log \left(\frac{1}{|P(i)|} \sum_{p \in P(i)} \frac{\exp(z_i \cdot z_p \,/\, \tau)}{\sum_{\alpha \in I - \{i\}} \exp(z_i \cdot z_\alpha \,/\, \tau)} \right) \tag{1}$$

The positive samples are usually chosen to be the samples that belong to the same class as the anchor. For our approach, we wish to contrast the image tokens that correspond to the objects for a given class and attribute (referring expression) to the ones that correspond to the same class but different attributes or to different classes altogether. To do this without explicit supervision for the image tokens $z_i \in I$ we leverage N, the number of ground truth counts in the image for the given RE and use it to assign pseudo-labels by dividing our image tokens into a positive (I^+) and negative (I^-) class. Given an image and RE tokens t, we get the masked RE tokens t_m by masking the token corresponding to the class label. Then we compute similarity scores y_i between all image tokens and the masked RE and assign the positive class to the N tokens with the highest similarity scores.

$$y_i = \frac{z_i \cdot t_m}{\|z_i\| \, \|t_m\|} \tag{2}$$

$$I^+ = \underset{i \in I}{\text{argtopN}}(y_i) \tag{3}$$

We also propose a modified version of the supervised contrastive loss, as the standard formulation also pushes samples from the negative class I^- closer together. In our case, this is undesirable since the negative class may contain samples from a diverse set of class labels, referring expressions, or even tokens corresponding to no class label. To address this, we modify the supervised contrastive objective to only use samples from the positive class as anchors. Based on this modification, our revised supervised contrastive loss can be formulated as follows:

$$\mathcal{L}^*_{\text{sup}} = \tag{4}$$
$$\underbrace{\sum_{i \in I^+}}_{\substack{\textbf{optimize} \\ \textbf{positives} \\ \textbf{only}}} -\log \left(\frac{1}{|I^+| - 1} \sum_{p \in I^+ - \{i\}} \frac{\exp(z_i \cdot z_p \,/\, \tau)}{\sum_{\alpha \in I - \{i\}} \exp(z_i \cdot z_\alpha \,/\, \tau)} \right)$$

We verify that the motivation for this modification is well-grounded and supported by demonstrating its effectiveness in our ablation study (Section 4.5).

Our final proposed model, named **C-REX** (**C**ontrastive learning for **R**eferring **Ex**pression Counting) combines the proposed modified detection losses, the L1 point center localization loss and the cross-entropy classification loss, and our proposed novel contrastive loss via a weighted sum:

$$\mathcal{L} = \lambda_{loc}\mathcal{L}_{loc} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_c \mathcal{L}^*_{\text{sup}} \tag{5}$$

3.3. Extension to class-agnostic counting

C-REX is designed to be general and versatile, enabling models to distinguish objects with subtle, context-dependent variations, such as those described by referring expressions. This allows us to adapt it to any task that requires multiple image-space object embeddings be contrasted and separated based on a set of distinguishing attributes. Following this principle, we introduce a simple

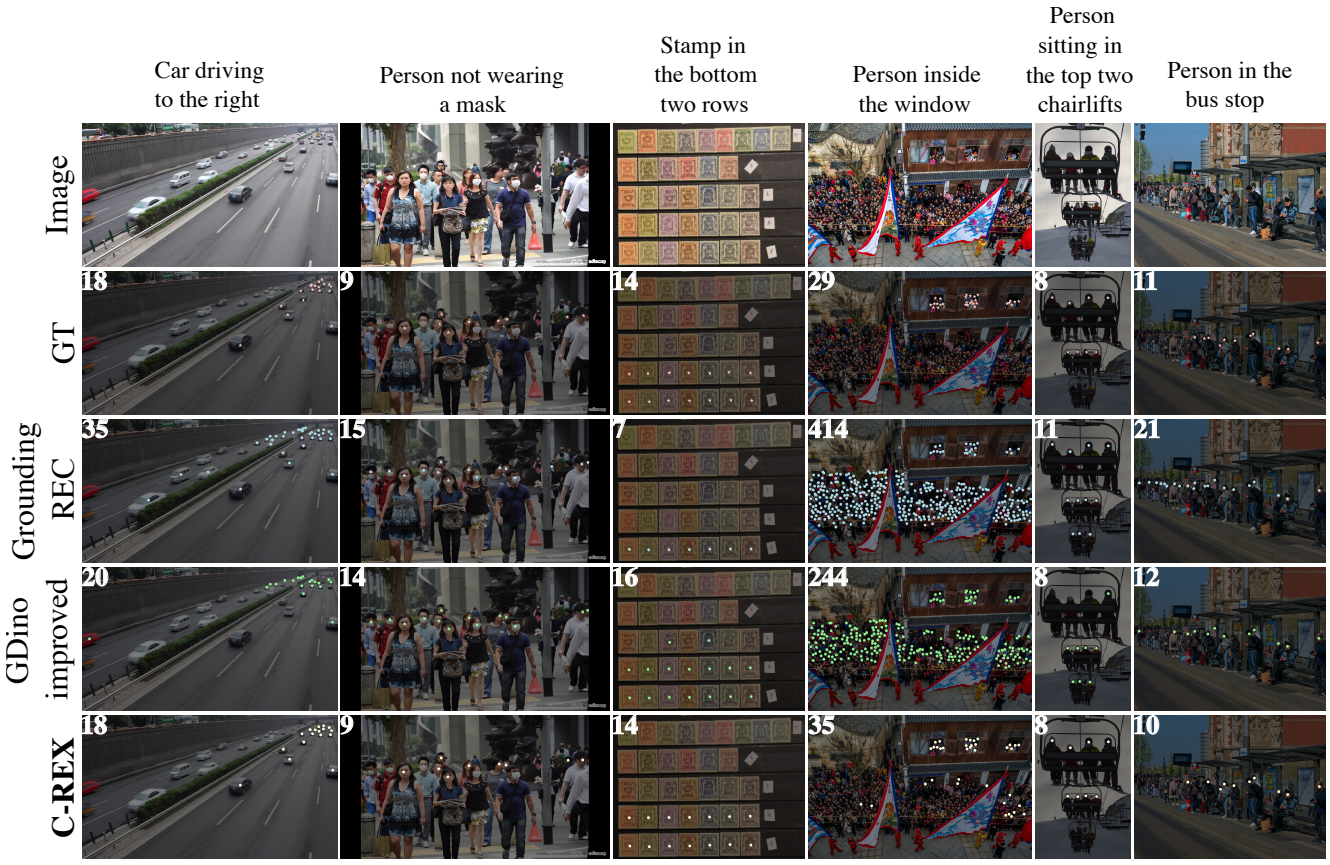

Figure 2. **Qualitative comparison** of referring expression counting (REC) results across different methods. The first row shows the input images, while the second row contains ground truth annotations. The third, fourth, and fifth rows display predictions from GroundingREC [12], our improved Grounding DINO baseline, and C-REX, respectively. Each column corresponds to a different referring expression. We observe that our method not only gets the most accurate counts, but it also counts the correct ground truth instances (i.e. the ones that were truly referred to by the given expression).

adaptation for class-agnostic counting, where positive samples are selected based on their similarity to the target class, rather than a referring expression. This adaptation is less critical in class-agnostic counting where models need to differentiate between objects of different classes, whose image embeddings are already different enough, so we do not expect to see the same amount of improvement as in REC. With this adaptation, our model achieves competitive performance in the class-agnostic counting task, outperforming previous text-based methods, as shown in Section 4, although the improvement is less significant, as was expected.

3.4. Advantages over alternate CL methods

Dai et al. [12] also tried to learn more discriminative features by contrasting image embeddings to text embeddings of "candidate" referring expressions. However, their approach has key limitations in the context of contrastive learning. Specifically, their contrastive loss operates on image-text pairs, which presents additional challenges due to the inherent misalignment between image and text embedding spaces [28]. In contrast, our method operates entirely within the image space, ensuring a more stable contrastive signal. CL in an image-text setting generally requires more data and larger-scale training to achieve robust representations [10, 40], whereas contrastive learning in the image space is more efficient [35, 60].

Additionally, a crucial factor in contrastive learning is the presence of a large batch size with numerous negative samples, as this increases the diversity and variability of samples included, thus improving the robustness of the learned representations [9, 17, 18, 25, 47]. However, [12] is limited in that regard, as the number of negative samples is constrained by the number of existing referring expressions for a single class in an image, typically fewer than four. Our contrastive loss overcomes this constraint by leveraging a significantly larger pool of negatives, typically in the hundreds, consisting of image embeddings not selected by our strategy [29]. These advantages are reflected in our ex-

perimental results (Section 4), where our model achieves state-of-the-art REC performance, substantially surpassing previous work in both MAE and RMSE.

4. Experiments

We train C-REX on the REC-8K training set, and then evaluate on its test split for the REC task, following Dai et al. [12]. We then train our model on the FSC-147 [43] training set only, and evaluate on the FSC-147 test set, and the CARPK test set, following the protocol of the state-of-the-art class-agnostic counting models [3, 36] to ensure a fair comparison for the class-agnostic counting task.

4.1. Implementation details

We use Grounding DINO [29] with a SWIN-T [31] image encoder and a BERT-base [14] text encoder, keeping them frozen and only finetuning its feature enhancer and cross-modality decoder. For the model output we select 900 tokens to make predictions following DETR [7] and use the thresholding method of Dai et al. [12] to select positive detections, setting a threshold of 0.30 for the CLS token and 0.36 for the rest of the text tokens. The predicted count is calculated as the number of predicted positive detections.

We train our models using AdamW with a learning rate of 1e-5 and a weight decay of 1e-4. For our loss calculation we select λ_{loc} to be 1, λ_{cls} to be 5 and λ_c to be 0.005.

4.2. Datasets & Metrics

REC-8K [12] consists of $8,011$ images, each paired with 2.13 referring expressions (REs) on average. In total, the dataset contains $17,122$ image-RE pairs, where each pair is annotated with an arbitrary number of ground-truth points pinpointing the target objects in the image.

FSC-147 [43] is a widely used object counting dataset, which consists of $6,135$ images spanning 147 classes, split in a non-overlapping manner. Each image is annotated with three visual exemplars.

CARPK [20] contains drone-captured images of cars typically located in parking lots. It has $1,448$ images, annotated with at least two bounding boxes as visual exemplars.

Metrics Following prior work [3, 12, 36, 55], we report the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to evaluate and compare performance. For REC we also report the F1 score following Dai et al. [12].

4.3. Results in REC

Quantitative Comparison We compare our proposed method, C-REX, against existing class-agnostic counting methods, including ZSC [55], TFOC [58], CounTX [2], CountGD [3], as well as GroundingREC [12], the only other method specifically designed for REC. Table 1 presents our experimental results, where we also compare with Grounding DINO [29], following Dai et al. [12], as well as our improved baseline. The latter surpasses GroundingREC in MAE and RMSE, though it falls slightly behind in F1 score. Noteably, C-REX outperforms all previous methods by a substantial margin, achieving the best results across all three metrics, becoming the sota model on REC-8K.

Method	Setting	Val set		Test set	
		MAE	RMSE	MAE	RMSE
FamNet [42]	VE	23.75	69.07	22.08	99.54
BMNet+ [44]	VE	15.74	58.53	14.62	91.83
CounTR [8]	VE	13.13	49.83	11.95	91.23
CACViT [52]	VE	10.63	37.95	9.13	48.96
DAVE [36]	VE	8.91	28.08	8.66	**32.36**
CountGD [3]	Both	**7.10**	**26.08**	**6.75**	43.65
ZSC [55]	Text	26.93	88.63	22.09	115.17
CounTX [2]	Text	17.10	65.61	15.88	106.29
CountGD [3]	Text	12.14	**47.51**	14.76	120.42
DAVE [36]	Text	15.48	52.57	14.90	103.42
GDino [29]	Text	10.32	55.54	10.82	104.00
GREC [12]	Text	10.06	58.62	10.12	107.19
GDino impr.	Text	**9.71**	55.11	10.73	103.79
C-REX	Text	10.19	57.14	**10.01**	**101.46**

Table 2. Comparison of results on the FSC-147 [43] dataset, using either text or visual exemplars as a guide to perform counting. GDino is short for GroundingDino, GREC is short for GroundingREC, and DGino impr. denotes our improved baseline. VE stands for visual exemplars, while *Both* stands for VE and text. All results from the original papers. Best results in bold.

Dataset	Method	Test	
		MAE ↓	RMSE ↓
CARPK	CLIP-count [23]	11.96	16.61
	CounTX [2]	8.13	10.87
	VLCounter [24]	6.46	8.68
	CountGD [3] †	3.83	5.41
	GroundingREC [12]	7.91	10.18
	GDino improved	**3.80**	**5.16**
	C-REX (ours)	4.21	6.12

Table 3. Comparison with state-of-the-art CAC methods, using text as guidance. All models are trained on FSC-147 [43]. We mark CountGD with † as it is the only model trained using both visual exemplars and text. CounTX is finetuned on CARPK [20].

Figure 3 presents the MAE and RMSE scores across different object count ranges on the test set, illustrating performance under varying object densities. Our proposed

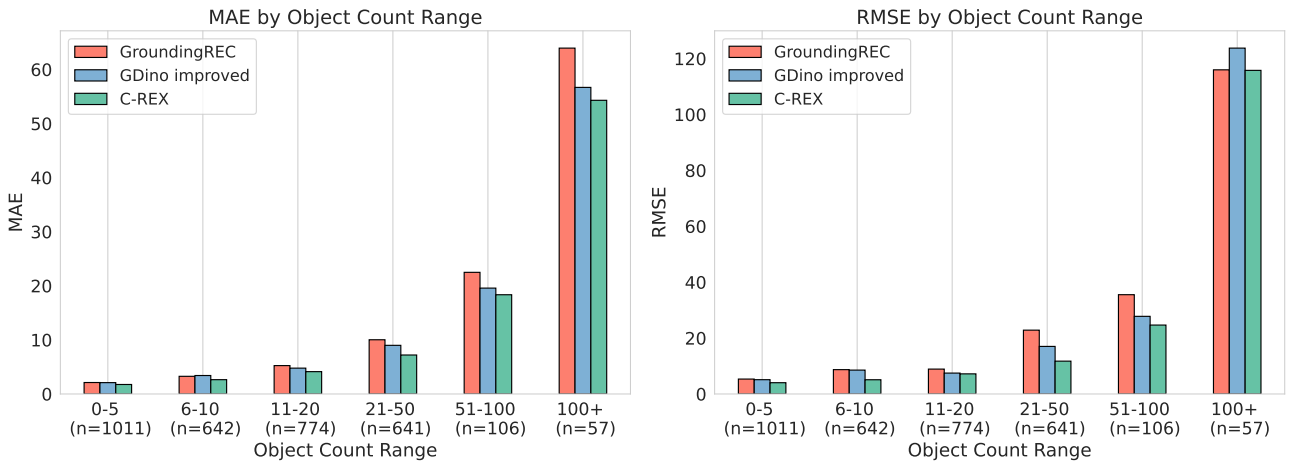

Figure 3. Quantitative comparison between GroundingREC, the improved GDino baseline and C-REX in the REC-8K test set by object count range. The number of samples for each bin is annotated below the bin's range as n. We observe that C-REX outperforms the two baseline models across all count ranges, with the results only being close for the RMSE higher count bin.

method, C-REX, achieves consistently lower MAE than GroundingREC and our improved Grounding DINO baseline across all object count ranges, with particularly noticeable improvements in mid-to-high count scenarios (21–100 objects). While the RMSE improvements for 100+ objects are more modest, C-REX still demonstrates comparable or superior performance, indicating the effectiveness of detecting object centroids instead of bounding boxes. Figure 4 presents the performance of the models across different attribute categories. C-REX generally outperforms both baselines, with the largest improvements observed in attributes like action, location, and color, where precise object differentiation is crucial (e.g., "car driving *left/right*"), showcasing the importance of our contrastive approach for disambiguating between items in those categories. Visual examples for these attribute categories are provided in the supplementary.

Qualitative Comparison Figure 2 presents qualitative comparisons between GroundingREC, our improved baseline, and C-REX on diverse REC scenarios. Each column corresponds to a different referring expression, ranging from object orientation ("car driving to the right"), to fine-grained attributes ("person not wearing a mask") and positional relationships ("stamp in the bottom two rows" or "person in the bus stop"). We observe that C-REX consistently provides more accurate counts, particularly in cases requiring precise attribute understanding and spatial awareness. For instance, in the "car driving to the right" scenario, C-REX correctly identifies the orientation of cars and accurately predicts the total count, whereas both baselines fail due to incorrect localization. Similarly, in the "stamp in the bottom two rows" example, C-REX correctly focuses on the relevant stamps, while other methods struggle with miscounting due to either restricting the count to one row or identifying objects from other rows. A similar issue arises in the "person in the bus stop" case, where the baselines fail to restrict counting to only those inside the bus stop, whereas C-REX demonstrates better selectivity and accuracy. We provide more examples of both high performing and failure cases for our model in the supplementary.

4.4. Results in Class-Agnostic Counting

We adapt C-REX for class-agnostic counting and evaluate it on the benchmark datasets FSC-147 [43] and CARPK [20]. Specifically, we trained C-REX on FSC-147 and then evaluated its generalization abilities on CARPK's test set. We present results for both datasets on Table 2 and Table 3 respectively. In FSC-147, C-REX outperforms all previous text-based methods on both MAE and RMSE on the test set. In CARPK, we notice that the more generic approach of the improved Grounding DINO baseline leads to the best overall performance, however C-REX is also competitive, performing better or on par with most previous methods.

4.5. Ablation Study

4.5.1. Supervised CL modification

On Table 4 we see that our proposed modification ($\mathcal{L}^*_{\text{sup}}$) to the supervised contrastive objective significantly improves performance compared to both the unmodified version and the baseline method. This highlights the importance of our decision to use only positive samples as anchors, confirming its crucial role in improving the model's effectiveness.

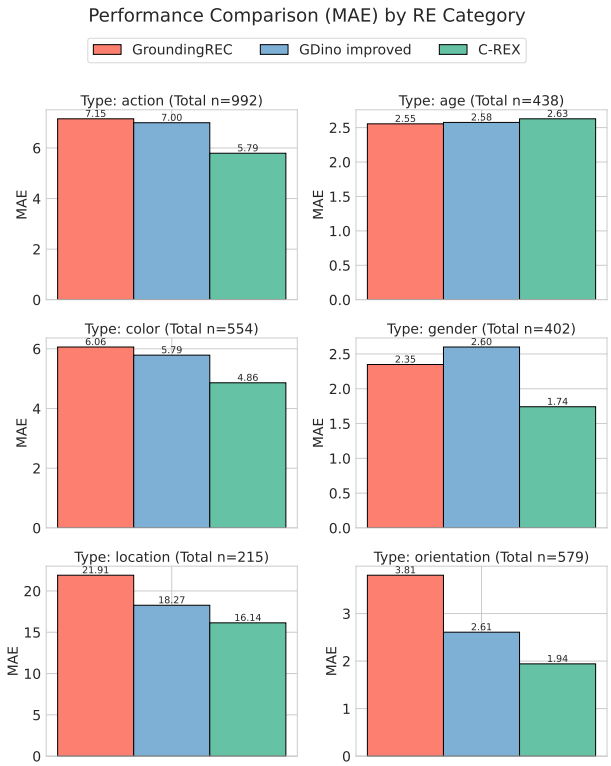

Figure 4. Quantitative comparison in terms of MAE between GroundingREC, our improved baseline and C-REX in the REC-8K test set for different RE categories. C-REX outperforms both baseline models across most categories, with the largest improvements shown in the *action*, *orientation* and *location* categories. We visualize categories with more than 30 samples.

4.5.2. Number of selections

We also verify that choosing the top N (where N corresponds to the ground truth count) samples to belong to the positive class outperforms two alternative strategies: (i) choosing a predefined small amount of high similarity samples that behave as pseudo-exemplars, and (ii) selecting slightly more than N samples to account for uncertainty in the selection process (Table 5). We discuss our choise of N further in the supplementary, showing that it is both theoretically justified and empirically supported.

4.5.3. Key baseline components

To validate the impact of re-purposing open-set detectors as object centroid predictors, we compare different variations of GroundingREC [12] and display the results in Table 6. First, we evaluate the full GroundingREC model, then we remove its feature fusion modules, and finally its CL module, leaving only the core detection-based counting pipeline which forms our baseline. The results, presented in Section 4.5, demonstrate that even without fusion modules or contrastive learning, our improved baseline remains competitive, highlighting the effectiveness of centroid-based detection.

Method	**Val set**			**Test set**		
	MAE	RMSE	F1	MAE	RMSE	F1
baseline	5.92	17.09	0.65	5.90	19.73	0.68
+ $\mathcal{L}_{\text{sup}}$	5.69	14.98	0.65	5.91	19.34	0.68
+ $\mathcal{L}^*_{\text{sup}}$	**4.86**	**13.60**	**0.68**	**5.06**	**17.53**	**0.70**

Table 4. Ablation study comparing the performance of C-REX on the validation and test set when using the proposed modified supervised contrastive loss ($\mathcal{L}^*_{\text{sup}}$) versus the typical supervised contrastive loss formulation ($\mathcal{L}_{\text{sup}}$). Our proposed modification leads to better performance across both datasets and metrics.

Number of Selections	**Val set**		**Test set**	
	MAE	RMSE	MAE	RMSE
top N	**4.86**	**13.60**	**5.06**	**17.53**
top 5	7.58	18.34	7.51	20.83
top $N + \sqrt{N}$	6.02	15.79	6.07	18.80
top $N + 2log_2(N)$	5.41	14.89	5.53	18.95

Table 5. Ablation study for the number of selections for our proposed contrastive loss. N refers to the ground truth object count.

Number of Selections	**Val set**		**Test set**	
	MAE	RMSE	MAE	RMSE
GDino	9.03	21.98	8.88	21.95
GREC [12]	6.80	18.13	6.50	19.79
GREC - feat.fusion	6.53	18.57	6.16	18.86
GDino impr.	**5.92**	**17.09**	**5.90**	**19.73**

Table 6. Ablation study for our proposed baseline. GREC is GroundingREC [12]. GREC - feat.fusion refer to GroundingREC without its feature fusion module. GDino impr. is our baseline.

5. Conclusion

In this work, we introduce C-REX, a novel contrastive learning framework designed to tackle Referring Expression Counting (REC) by improving detection-based counting models' capabilities in distinguishing visually similar objects with different referring expressions. By operating entirely within the image space, C-REX eliminates the misalignment issues inherent in image-text contrastive learning, ensuring a more stable contrastive signal and leveraging a significantly larger pool of negative samples for improved robustness. Additionally, we adapt C-REX for class-agnostic counting, and explain that it is general and versatile; being able to be adapted for other visual tasks that require discriminating between visually similar but contextually distinct objects like referring expression segmentation or detection. To support our approach, we analyzed sota

detection-based counting methods and identified centroid-based detection as a key factor behind their success, using it to design an improved detection-based baseline. Our experiments validate the effectiveness of C-REX, that achieves sota performance in REC with over 22% improvement in MAE and more than 10% in RMSE, while also performing strongly in class-agnostic counting. These findings highlight the potential of CL within the image space for broader object counting and vision-language tasks, opening avenues for future research in fine-grained visual understanding.

References

[1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):872–881, 2021. 2

[2] Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specified object counting. *ArXiv*, abs/2306.01851, 2023. 2, 4, 6

[3] N. Amini-Naieni, T. Han, and A. Zisserman. Countgd: Multi-modal open-world counting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3, 4, 6

[4] Carlos Arteta, Victor Lempitsky, J. Alison Noble, and Andrew Zisserman. Interactive object counting. In *Computer Vision – ECCV 2014*, pages 504–518, Cham, 2014. Springer International Publishing. 3

[5] C S Asha and A V Narasimhadhan. Vehicle counting for traffic management system using yolo and correlation filter. In *2018 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–6, 2018. 2

[6] Varun Belagali, Srikar Yellapragada, Alexandros Graikos, Saarthak Kapse, Zilinghan Li, Tarak Nath Nandi, Ravi K Madduri, Prateek Prasanna, Joel Saltz, and Dimitris Samaras. Gen-sis: Generative self-augmentation improves self-supervised learning. *arXiv preprint arXiv:2412.01672*, 2024. 12

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. 6

[8] Liu Chang, Zhong Yujie, Zisserman Andrew, and Xie Weidi. Countr: Transformer-based generalised visual counting. In *British Machine Vision Conference (BMVC)*, 2022. 2, 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 3, 5

[10] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023. 3, 5

[11] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *CVPR*, 2019. 3

[12] Siyang Dai, Jun Liu, and Ngai-Man Cheung. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16985–16995, 2024. 2, 3, 4, 5, 6, 8, 12

[13] Zhe Dai, Huansheng Song, Xuan Wang, Yong Fang, Xu Yun, Zhaoyang Zhang, and Huaiyu Li. Video-based vehicle counting framework. *IEEE Access*, 7:64460–64470, 2019. 2

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 6

[15] Qiyan Fu, Weidong Min, Weixiang Sheng, and Chunjiang Peng. Counting dense object of multiple types based on feature enhancement. *Frontiers in Neurorobotics*, 18, 2024. 2, 3

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3, 5

[18] Olivier J Henaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020. 3, 5

[19] Michael Hobley and Victor Prisacariu. Abc easy as 123: A blind counter for exemplar-free multi-class class-agnostic counting. In *European Conference on Computer Vision*, pages 304–319. Springer, 2024. 2

[20] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H. Hsu. Drone-based object counting by spatially regularized regional proposal network. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4165–4173, 2017. 6, 7

[21] Yifeng Huang, Duc Duy Nguyen, Lam Nguyen, Cuong Pham, and Minh Hoai. Count what you want: exemplar identification and few-shot counting of human actions in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10057–10065, 2024. 2

[22] Xiaofei Hui, Qian Wu, Hossein Rahmani, and Jun Liu. Class-agnostic object counting with text-to-image diffusion model. In *Computer Vision – ECCV 2024*, pages 1–18. Springer Nature Switzerland, 2025. 2

[23] Ruixia Jiang, Lin Liu, and Changan Chen. Clip-count: Towards text-guided zero-shot object counting. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 6

[24] 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*, pages 2714–2722, 2024. 2, 6

[25] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2, 3, 5

[26] D. Kong, D. Gray, and Hai Tao. A viewpoint invariant approach for crowd counting. In *18th International Conference on Pattern Recognition (ICPR'06)*, pages 1187–1190, 2006. 3

[27] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2010. 3

[28] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 5

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *European Conference on Computer Vision*, 2024. 2, 3, 4, 5, 6

[30] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 6

[32] Erika Lu, Weidi Xie, and Andrew Zisserman. Class-agnostic counting. In *Computer Vision – ACCV 2018*, pages 669–684, Cham, 2019. Springer International Publishing. 2

[33] E-Ro Nguyen, Hieu Le, Dimitris Samaras, and Michael Ryoo. Instance-aware generalized referring expression segmentation. *arXiv preprint arXiv:2411.15087*, 2024. 3

[34] Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, page 348–365, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 3, 5, 12

[36] Jer Pelhan, Alan Lukevzivc, Vitjan Zavrtanik, and Matej Kristan. Dave – a detect-and-verify paradigm for low-shot counting. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23293–23302, 2024. 3, 6

[37] Jer Pelhan, Alan Lukezic, Vitjan Zavrtanik, and Matej Kristan. A novel unified architecture for low-shot counting by detection and segmentation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[38] Roland Perko, Manfred Klopschitz, Alexander Almer, and Peter M. Roth. Critical aspects of person counting and density estimation. *Journal of Imaging*, 7, 2021. 2, 3

[39] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020. 3

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3, 5

[41] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[42] 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 (CVPR)*, 2021. 2, 6

[43] 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 (CVPR)*, 2021. 6, 7

[44] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and ZHIGUO CAO. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9519–9528, 2022. 6

[45] Wei Su, Peihan Miao, Huanzhang Dou, and Xi Li. Scanformer: Referring expression comprehension by iteratively scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13449–13458, 2024. 3

[46] Yuejiao Su, Yi Wang, Lei Yao, and Lap-Pui Chau. Few-shot class-agnostic counting with occlusion augmentation and localization. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2024. 2

[47] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 3, 5

[48] Jia Wan, Ziquan Liu, and Antoni B. Chan. A generalized loss function for crowd counting and localization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1974–1983, 2021. 2

[49] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. In *Advances in Neural Information Processing Systems*, 2020. 2

[50] Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 12998–13008, 2024. 3

[51] Yutian Wang, Bin Yang, Xi Wang, Chao Liang, and Jun Chen. Satcount: A scale-aware transformer-based class-agnostic counting framework. *Neural Networks*, 172: 106126, 2024. 2

[52] Zhicheng Wang, Liwen Xiao, Zhiguo Cao, and Hao Lu. Vision transformer off-the-shelf: A surprising baseline for few-shot class-agnostic counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 6

[53] Changli Wu, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, Rongrong Ji, et al. Rg-san: Rule-guided spatial awareness network for end-to-end 3d referring expression segmentation. *Advances in Neural Information Processing Systems*, 37:110972–110999, 2025. 3

[54] Jingyi Xu, Hieu M. Le, and Dimitris Samaras. Learning from pseudo-labeled segmentation for multi-class object counting. *WACV 2025*, abs/2307.07677, 2023. 2

[55] Jingyi Xu, Hieu M. Le, and Dimitris Samaras. Zero-shot object counting with language-vision models. *ArXiv*, abs/2309.13097, 2023. 4, 6

[56] Yuanwu Xu, Feifan Song, and Haofeng Zhang. Learning spatial similarity distribution for few-shot object counting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024. 2

[57] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 12

[58] Mengmi Zhang Zenglin Shi, Ying Sun. Training-free object counting with prompts. In *WACV*, 2024. 2, 4, 6

[59] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 12

[60] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5, 12

[61] Qi Zhang and Antoni B. Chan. Calibration-free multi-view crowd counting. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, page 227–244, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[62] Huang Zhizhong, Dai Mingliang, Zhang Yi, Zhang Junping, and Shan Hongming. Point, segment and count: A generalized framework for object counting. In *CVPR*, 2024. 2

Improving Contrastive Learning for Referring Expression Counting

Supplementary Material

We organize the supplementary material as follows:

6. Discussion on choice of number of selections

Intuitively choosing the the top N (where N is the ground truth count) samples based on their similarity to the referring expression makes sense, as visually similar objects with incorrect referring expressions often lie near the decision boundary. In the context of contrastive learning, it is crucial to target excluding those samples in order to build more discriminative representations for samples with fine-grained differences. We verify that when selecting with this strategy, approximately 80% of selected samples correspond to correct instances. This shows that if there are N correct samples and $\tilde{N}$ additional objects that are visually similar but correspond to different referring expressions, the selection process still favors correct samples. Unless $\tilde{N} << N$, we guarantee that most of the correct samples will be pushed away from most of the "confusing" ones. Moreover, some degree of label uncertainty is not detrimental, as prior research has shown that controlled label noise can even improve model robustness and generalization [6, 57, 59]. In fact, incorporating some degree of label uncertainty is a relatively common practice in self-supervised learning. For example, in DINO training, random local crops often exclude key objects or contain only background, yet the model still learns meaningful representations [35, 60].

7. Details on metrics

We use MAE, RMSE and F1 as metrics. Given n samples where c_i represents the ground truth counts and $\hat{c}_i$ represents the predicted counts for each sample i, the MAE and RMSE are calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|c_i - \hat{c}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(c_i - \hat{c}_i)^2} \quad (6)$$

In the REC task specifically, images typically contain multiple instances of a given class, each associated with different referring expressions. This setup poses a significant challenge for detection-based counting models, which may miss correct instances or generate false positives by misidentifying objects of the same class that belong to different referring expressions. When this occurs, MAE and RMSE can be misleadingly low, failing to capture the true counting performance. To address this, we also report the F1 score, following Dai et al. [12], which ensures that our predicted count not only matches the total number of instances but also correctly identifies them.

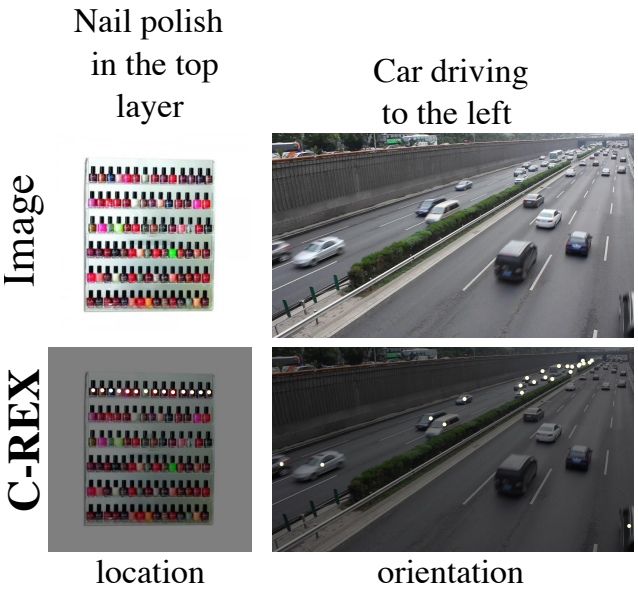

Figure 5. Some qualitative examples from the "location" and "orientation" attribute categories, for which our model vastly outperforms previous works. Here we can observe that our novel contrastive learning approach allows the model to disambiguate between fine-grained spatial attributes, only selecting instances from items in the "top" layer for the first image and cars driving to the "left" in the second image.

8. Qualitative Results from well performing attribute categories

Figure 5 showcases the model's localization and orientation capabilities. Specifically, the first column of Figure 5 presents a cardboard with multiple rows of nail polishes, with the target referring expression prompting the model to count the ones located in the top shelf. We observe that the model succesfully locates and subsequently counts all the nail polishes found in the top shelf, and none found in the rest of the shelves. In the second column, we observe the model's good orientation capabilities, as it manages to correctly count the cars driving on the right side of the road, without misidentifying cars on the left side.

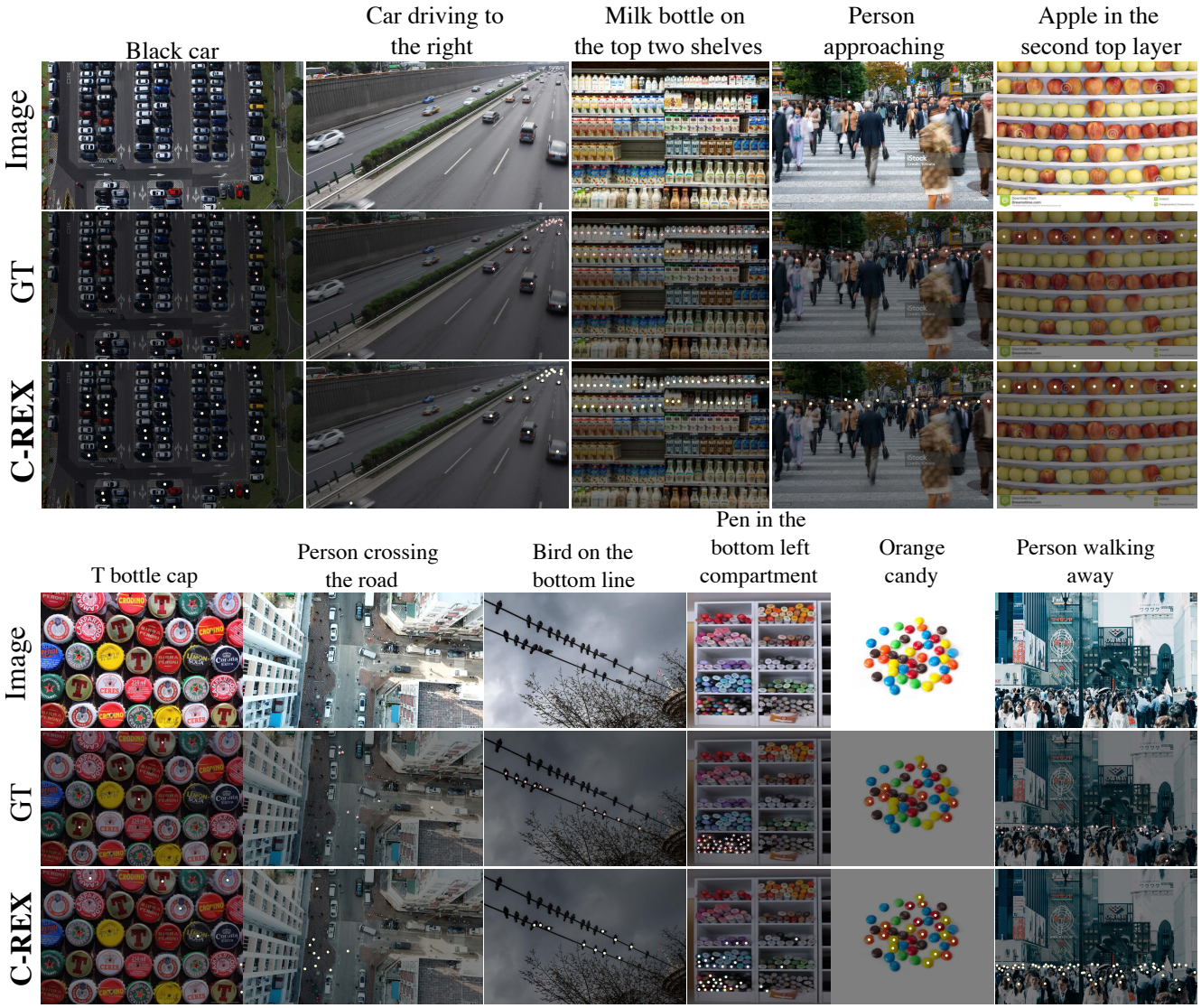

Figure 6. **Qualitative Comparison** of referring expression counting (REC) results where our proposed method C-REX achieves good performance. The first row shows the input images, while the second row contains ground truth annotations. The third row displays C-REX predictions. Each column corresponds to a different referring expression.

Figure 7. **Qualitative Comparison** of referring expression counting (REC) results where our proposed method C-REX yields poor performance. In the same notion, the first row shows the input images, while the second row contains ground truth annotations. The third row displays C-REX predictions. Each column corresponds to a different referring expression.

9. Additional Qualitative Results

In Figure 6, we present a series of examples where the proposed method achieved good and poor performance respectively. The former are presented in the top sub-figure, while the latter are found in the bottom sub-figure. For instance, in the first column of the top sub-figure, our proposed model successfully identifies all the black cars, while it does not count any differently colored car instances. Another notable example is the one on the fourth column ("Apple in the second top layer"), where the model correctly counts and locates the right row of apples, demonstrating its good performance on positional relationship attributes.

In contrast, there are also cases where C-REX fails to correctly count the target instances. Such cases are presented in the bottom sub-figure of Figure 6. For example, in the first column, where the target objects are defined as the bottle caps with the letter "T" on them, we notice that the model fails to identify, and thus correctly count, the vast majority of them. Moreover, in the third column, where the target referring expression prompts us to count the amount

of birds in the bottom line of the cables, we notice that the model miscounts the ones that are actually located in the bottom row, while it also counts three that lie in the top row.