Q-VG: TOWARDS QUANTITATIVE REASONING FOR WEAKLY SUPERVISED VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly supervised visual grounding (WSVG) aims to locate objects referred by natural language queries, without region-level human annotations. However, existing methods struggle to ensure the correct number of targets, a common requirement in compositional reasoning. In this paper, we propose a count-aware WSVG framework, Q-VG, for quantitative reasoning and grounding under weak supervision. Specifically, Q-VG incorporates a novel Referring-Object Counting (RECO) module to estimate the number of objects referred to by the input expression. It enhances quantitative awareness prior to grounding and determines the number of bounding boxes to regress. Besides, we also propose a novel weakly supervised objective, which encourages multiple positive objects to align with the given referring phrase. To further evaluate the ability of quantitative reasoning, we introduce QGround-Bench, a quantitative grounding benchmark. It comprises three subsets, Exact-QG, Gen-QG and Comp-QG, which progressively evaluate exact quantitative constraints, generalized quantitative expressions, and quantitative compositional reasoning. Experiments show that Q-VG achieves competitive performance on both the proposed and classical benchmarks. The source code and dataset will be available on GitHub after double-blind phase.

1 INTRODUCTION

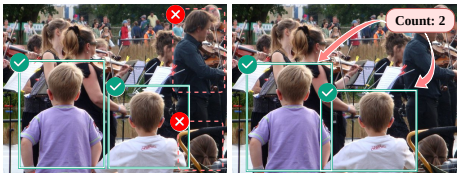
Weakly Supervised Visual Grounding (WSVG) aims to localize objects described by language queries without region-level annotations during training. This learning process reduces the reliance on expensive region-level annotations. Because grounding bridges vision and language, it has been integrated into various downstream applications, such as MLLM reasoning (Wang et al., 2026a), grounded VQA (Wang et al., 2025) and GUI agent (Li et al., 2025; Tang et al., 2026).

Previous WSVG studies learn visual-language alignment through auxiliary tasks, and produce heatmaps determining positions (He et al., 2025; Lin et al., 2026a). To regress target bounding box directly, some works generate proposals with pretrained detectors and then select the most relevant region (Kuang et al., 2025; Chen et al., 2025). Despite their progress, most approaches do not explicitly model how many objects should be returned, resulting in missed or redundant localization.

Here, two questions arise for solving these problems. **Firstly**, how to determine the number of referred objects? As shown in Figure 1, for the phrase "some boys onlooking during a music performance", the quantifier "some" and the plural form "boys" only provide coarse quantity cues without specifying the exact number of targets. Visual instances further help determine the actual set of referred boys. Thus, resolving the target quantity requires a joint understanding of linguistic cues and visual instances, making quantity inference a fundamental challenge in multi-object grounding.

Secondly, how to rethink the evaluation of multi-object visual grounding? Existing benchmarks mainly evaluate whether the final referent is correctly localized, while overlooking other entities involved in compositional reasoning. However, correctly localizing the final referent does not necessarily reflect whether the model understands the quantitative cues and entity roles in the expression. For example, grounding "the man next to two dogs" requires not only locating the man, but also understanding the quantity and role of the dogs. Therefore, multi-object evaluation should consider both the quantity and roles of relevant entities, and distinguish errors in category recognition from those in quantity understanding. To this end, we design a difficulty-aware quantitative grounding benchmark for systematically evaluating multi-entity grounding and quantity reasoning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Previous WSVG method (b) Our method

Figure 1: WSVG localization results. (a) produces redundant boxes, whereas (b) uses quantitative reasoning to select the referred objects.

Dataset	Target & Context	Multi-target	Quantity-specific Metrics	Difficulty Levels
Flickr30K Entities	✓	✓	✗	✗
ReferIt	✗	✗	✗	✗
RefCOCO	✗	✗	✗	✗
gRefCOCO	✗	✓	✗	✗
QGround-Bench	✓	✓	✓	✓

Table 1: Comparison of visual grounding datasets. QGround-Bench supports target–context entities, multi-target examples, quantity-specific metrics, and difficulty levels.

To solve these problems, we propose Q-VG, a count-aware WSVG framework for quantitative reasoning and grounding. Specifically, we utilize YOLO's anchor features to represent regional visual evidence, and choose the target ones based on anchor-text matching process. To ensure the quantity of referred objects, we design a Referring-Object Counting (RECO) module. This module defines phrase-related slots and counts the number of target objects. To encourage count-consistent positive anchors to align with the query phrase, we optimize the network using a weakly supervised objective. It includes multi-positive contrastive loss, reconstruction loss, and counting loss. Specifically, multi-positive contrastive loss builds one-to-many relation between a query phrase and target objects. Reconstruction loss preserves object-level visual details, while counting loss uses explicit quantitative cues as supervision.

To validate Q-VG, we conduct experiments on Flickr30K Entities (Plummer et al., 2015) and gRefCOCO (Ding et al., 2026). We also construct a novel benchmark dataset, QGround-Bench, to evaluate the ability of quantitative reasoning and grounding. Unlike existing datasets, it jointly annotates target and context entities, supports multi-target grounding, and organizes quantity-specific examples by difficulty, as shown in Table 1. QGround-Bench consists of three subsets: Exact-QG, Gen-QG and Comp-QG. They evaluate exact quantitative constraints, generalized quantitative expressions, and quantitative compositional reasoning, respectively. The experimental results show that Q-VG consistently outperforms existing WSVG methods in quantitative grounding.

Our main contributions are summarized as follows.

- We propose a count-aware WSVG framework, Q-VG, for quantitative reasoning and grounding. Q-VG counts the referred objects, and encourages multiple positive objects to align with the given referring phrase by a weakly supervised objective.
- We present QGround-Bench, a novel grounding dataset designed to assess the quantitative reasoning. Moreover, it progressively evaluates quantity-related exact constraints, generalized expressions, and compositional reasoning.
- We conduct extensive experiments on two classical datasets and QGround-Bench to demonstrate the effectiveness of our proposed Q-VG.

2 Methodology

Given an image I and a phrase P, we aim to generate bounding boxes $\mathcal{B}$ that localize the referred objects in the image. The overall framework of our Q-VG is illustrated in Figure 2.

2.1 Visual and Textual Encoders

In this section, we introduce three encoders in our Q-VG. We firstly extract the image feature $f_v \in \mathbb{R}^{H\times W\times D}$ using the encoder of pretrained YOLO (Redmon & Farhadi, 2018), where D denotes the feature dimension. The feature map contains $H \times W$ anchor features $f_a \in \mathbb{R}^D$. Based on them, the detection head outputs confidence scores and coordinates of bounding boxes. Also, we use DINOv3 (Siméoni et al., 2025) to extract dense semantic visual features as f_d, and a GloVe-based textual encoder to obtain token features $F_p \in \mathbb{R}^{L\times D}$ and the pooled representation $f_p \in \mathbb{R}^D$.

2.2 Count-Aware Anchor Selection

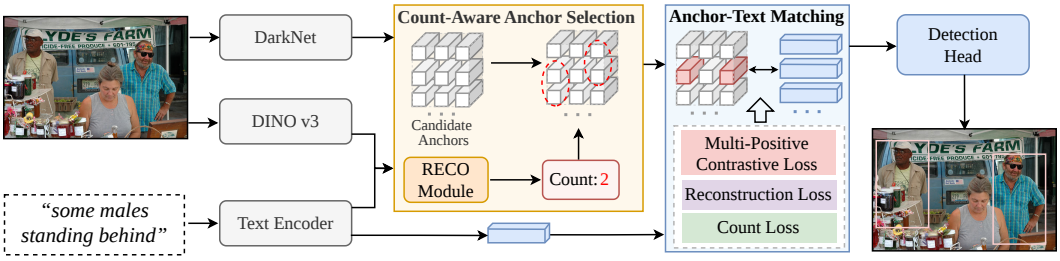


Figure 2: The overview of our proposed Q-VG. The image and phrase are first encoded by visual and text encoders. Then, the RECO module calculates the referring object counts. After that, Q-VG filters anchors not including targets and returns the matching ones for bounding box prediction.

In this section, we introduce the anchor selection process. Similar to contrastive pretraining (Radford et al., 2021), we learn vision-language alignment via image feature and textual embedding pairs. However, one quantitative phrase may refer to multiple objects, requiring the selection of multiple anchors. Thus, we should design the RECO module to count the targets and make sure the number of final box outputs.

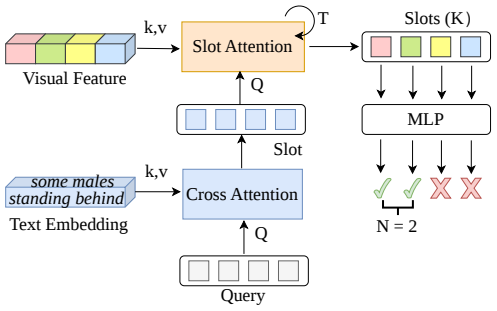


Figure 3: The architecture of RECO module.

RECO Module. The architecture of the module is shown in Figure 3. To effectively train the RECO module, we formulate the counting task as object existence prediction. Our goal is to make a distinction for multiple objects, and determine whether it is associated with the phrase or not. Thus, the target amount could be predicted as that of phrase-related ones. We firstly initialize the parameters of slots as $\mathrm{S} \in \mathbb{R}^{K\times D}$, which are defined as learnable tokens (Locatello et al., 2020). Here, K represents the total number of slots available to capture all potential objects. To enrich these slots as phrase-related ones, we utilize cross-attention to fuse the textual embedding f_p and these slots. Next, we update the output S^0 in an iterative design. These slots eventually become K object embeddings. Let $\phi : \mathbb{R}^{H\times W\times D} \mapsto \mathbb{R}^{K\times D}$ be this slot attention process. The object embeddings are obtained by iteratively updating initial phrase-related slots $\mathrm{S}^0 \in \mathbb{R}^{K\times D}$ T-times as follows,

$$\mathrm{S}^0 = \mathrm{LN}\left(\mathrm{S} + \mathrm{softmax}\left(\frac{q_s(\mathrm{S})k_p(F_p)^\top}{\sqrt{D_h}}\right)v_p(F_p)W_p\right), \tag{1}$$

where $q_s(\cdot)$, $k_p(\cdot)$, and $v_p(\cdot)$ are linear projections, and W_p maps the output to dimension D.

$$\begin{aligned}\Phi\left(f_d\right) &:= \mathrm{S}^T, \text{ where} \\ \mathrm{S}^t &= \mathrm{Agg}\left(f_d; \mathrm{S}^{t-1}\right),\end{aligned} \tag{2}$$

and $\mathrm{Agg}(\cdot)$ denote the aggregation process. To let each slot aggregate features of distinct objects, this process utilizes slots of previous iteration S^{t-1} with the DINO features f_d as inputs. Then, the attention map $A^{slot} \in \mathrm{R}^{H\times W\times K}$ between f_d and S^{t-1} is computed as follows,

$$A^{\text{slot}}_{i,j} = \frac{e^{M_{i,j}}}{\sum_{k=1}^{K} e^{M_{i,k}}}, M = \frac{k\left(f_d\right) q\left(\mathbf{S}^{t-1}\right)^\top}{\sqrt{D_h}}, \tag{3}$$

where $q(\cdot)$, $k(\cdot)$, and $v(\cdot)$ are linear projection layers with layer normalization that map inputs to dimension D_h. Note that this aggregation process encourages slots to compete for aggregating the image features. This competitive aggregation encourages different slots to capture distinct object-level evidence. Given the attention A^{slot}, the slots are updated as follows,

$$\begin{aligned}\mathrm{S}^t &= \mathrm{Agg}(f_d; \mathrm{S}^{t-1}) \\ &= \mathrm{normalize}\Big((A^{\text{slot}})^\top v(f_d)W + \mathrm{S}^{t-1}\Big),\end{aligned} \tag{4}$$

where $W \in \mathrm{R}^{D_h \times D}$ is a linear projection.

Finally, each slot $\{\mathrm{S}_k^T, k \in 1, ..., K\}$ produces a prediction value $p_k = \sigma(\mathrm{MLP}(\mathrm{S}_k^T))$ via an MLP followed by a sigmoid function σ. This value represents the existence probability of a target object. Here, the total predicted count N is the summation of these probabilities.

Anchor Selection. To select high-quality anchors, we retain the anchors in the last convolution feature map. Then, we filter these anchors by their confidence scores and constrain the number of selected anchors. To select positive examples for the training objectives, we compute the similarities between these anchors $\mathcal{A} = \{a_i\}$ and the phrase P. For each anchor $a \in \mathcal{A}$, we measure the spatial consistency between the anchor and the saliency region in A^{slot} as follows,

$$\mathcal{S}(a \mid I, P) = \mathrm{sim}(f_a, f_p) + \mathrm{AvgPool}_{\Omega_a}(\bar{A}), \tag{5}$$

where $\mathrm{sim}(\cdot)$ computes the similarity score via a lightweight linear projection. AvgPool computes the mean attention within the anchor region Ω_a. Here, $\bar{A} \in \mathbb{R}^{H \times W}$ is a scalar saliency map obtained by aggregating the K slot-attention channels according to their predicted existence probabilities. Finally, the set of positive anchors $\mathcal{A}^*$ is selected as follows,

$$\mathcal{A}^* = \underset{a \in \mathcal{A}}{\text{Top-}\hat{N}} \{\mathcal{S}(a \mid I, P)\}_{a \in \mathcal{A}}, \tag{6}$$

where $\mathcal{A}$ is the set of candidate anchors. Here, $\hat{N}$ is the discretized prediction of the RECO module.

2.3 Anchor-Text Matching

In this section, we detail our training objectives and the inference stage.

Multi-Positive Contrastive Loss. In WSVG, the phrase and target objects in an image are usually matched one-to-many. For each image i, we embed the phrase P_i and each anchor with lightweight projections as $f_p^{(i)}$ and $f_a^{(i)}$. Specifically, the loss increases the aggregate probability mass of the selected positive-anchor set $\mathcal{A}_i^*$, while treating remaining anchors in $\mathcal{A}_i \backslash \mathcal{A}_i^\star$ and mismatched image-phrase pairs as negatives. Thus, we define the inter- and intra-image contrastive losses as follows,

$$\mathcal{L}_{\text{mcl}} = \frac{1}{B_+} \sum_{i \in \mathcal{I}_+} \left[-\log \frac{\sum\limits_{a \in \mathcal{A}_i^\star} \exp\Big(\mathrm{sim}\big(f_p^{(i)}, f_a^{(i)}\big)/\tau\Big)}{\sum\limits_{a \in \mathcal{A}_i} \exp\Big(\mathrm{sim}\big(f_p^{(i)}, f_a^{(i)}\big)/\tau\Big) \; + \sum\limits_{\substack{j=1 \\ j \neq i}}^{B} \sum\limits_{a \in \mathcal{A}_j} \exp\Big(\mathrm{sim}\big(f_p^{(i)}, f_a^{(j)}\big)/\tau\Big)} \right], \tag{7}$$

where B denotes the batch size, τ is a temperature parameter, $\mathcal{I}_+ = \{i \mid |\mathcal{A}_i^\star| > 0\}$ denotes the set of non-empty samples in the mini-batch, and $B_+ = |\mathcal{I}_+|$.

Reconstruction Loss. In WSVG, the slot attention module needs to preserve visual information during its iterative refinement. Thus, we provide consistent supervision during reconstruction process, and design a lightweight decoder to map the slot features S^T back into the visual feature space. It then yields the reconstructed feature $\hat{\mathbf{f}}_d \in \mathbb{R}^{H \times W \times D}$. To better align reconstruction with phrase-related visual feature and slot, we utilize the Mean Squared Error and design the loss as follows,

$$\mathcal{L}_{rec} = \frac{1}{HWD} \sum_{i=1}^{H} \sum_{j=1}^{W} \left\| \delta(f_p) \odot \mathbf{f}_d^{(i,j)} - \hat{\mathbf{f}}_d^{(i,j)} \right\|_2^2 \tag{8}$$

where δ denotes a projection head via MLP, and $\odot$ means Hadamard product. This auxiliary objective regularizes the phrase-conditioned slots to preserve visual details during iterative refinement.

Counting Loss. Phrases in weakly supervised grounding often contain explicit numerals (e.g., "three", "four"), vague quantifiers (e.g., "some", "several"), or no quantitative information at all. To leverage these cues, we design a counting loss $\mathcal{L}_{cnt}$ as follows,

$$\mathcal{L}_{cnt} = \frac{\sum_{n=1}^{B} v_n \left([C_{n,\text{low}} - N_n]_+^2 + \mathbb{I}_n \left[N_n - C_{n,\text{high}} \right]_+^2 \right)}{\sum_{n=1}^{B} v_n + \epsilon}, \tag{9}$$

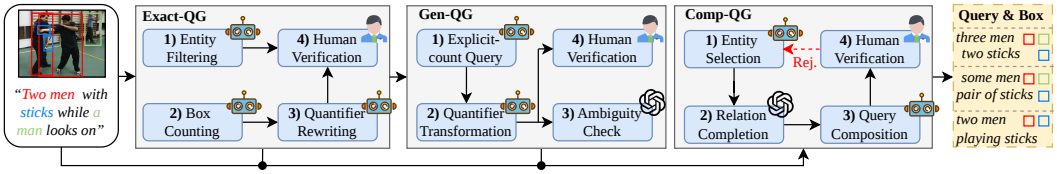


Figure 4: The pipeline of data construction for three QGround-Bench subsets, including Exact-QG, Gen-QG, and Comp-QG.

where B is the batch size, $[\cdot]_+ = \max(0, \cdot)$, and N_n is the predicted count for the n-th sample. $C_{n,\text{low}}$ and $C_{n,\text{high}}$ denote its lower and upper count constraints. The validity mask v_n excludes expressions without reliable quantitative supervision, whereas $\mathbb{I}_n$ activates the upper-bound term when a finite upper bound is available. For our network, we utilize the total loss as follows,

$$L = L_{mcl} + \lambda_r L_{rec} + \lambda_c L_{cnt} \tag{10}$$

Here, λ_r and λ_c denote the weights for the corresponding losses.

Inference. During the inference stage, our model aims to predict the Top-N most relevant bounding boxes given the input image I and phrase P. Specifically, we first identify the best-matched anchors A^* following Eq. 6. Then, the final bounding boxes $\mathcal{B}$ are regressed from the selected anchors $a^* \in \mathcal{A}^*$ via a prediction head as follows,

$$b = \text{Head}(a^*), \quad \forall a^* \in \mathcal{A}^*, \tag{11}$$

where $\mathcal{B} = \{b_1, b_2, \ldots, b_N\}$ constitutes the final set of grounded results. Note that $N = 0$ denotes an empty set $\mathcal{B}$, corresponding to the case where no target object is present.

3 QGround-Bench

3.1 Dataset

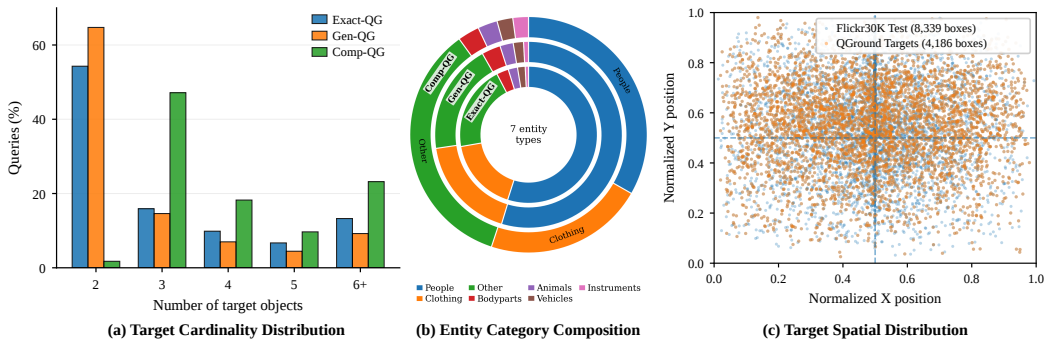
In order to construct a quantitative grounding benchmark, we require images with object bounding box annotations and expressions involving explicit quantities, generalized quantifiers, and compositional relations. Therefore, we build QGround-Bench upon the Flickr30K Entities test split, which contains 1,000 images paired with long captions describing multiple objects and their visual contexts. The data construction pipeline is illustrated in Figure 4.

To comprehensively evaluate different levels of quantitative understanding, QGround-Bench is divided into three subsets: Exact-QG, Gen-QG, and Comp-QG. Each sample contains multiple referred objects with corresponding descriptions, covering exact quantities, generalized quantitative expressions, and compositional quantities. The basic statistics and distribution characteristics of QGround-Bench are summarized in Figure 5.

Exact-QG. Our primary goal is to examine whether models can ground all referred objects according to an explicit quantity. To construct such samples, we first filter entity phrases associated with multiple bounding boxes and count the number of corresponding instances. We then rewrite the original phrase by introducing an exact cardinal number corresponding to the box count. Note that an image may contain additional objects of the same category that are not referred to in the caption, which can introduce ambiguity. Therefore, we manually remove such ambiguous samples, as well as samples whose quantities are inconsistent with the annotated bounding boxes.

Gen-QG. This subset aims to examine quantitative grounding beyond explicit cardinal numbers. Starting from the queries in Exact-QG, we replace explicit counts with generalized or lexicalized quantitative expressions while preserving the referred object set. For example, "three men" can be transformed into "some men", while "two sticks" can be expressed as "a pair of sticks". Since such transformations may introduce semantic ambiguity, we further use GPT-5.6[1] to check whether the transformed phrase uniquely corresponds to the annotated entities. The remaining samples are manually verified to ensure the consistency between the generated phrase and its referred objects.

[1]https://chatgpt.com/



(a) Target Cardinality Distribution **(b) Entity Category Composition** **(c) Target Spatial Distribution**

Figure 5: Statistics of QGround-Bench. (a) Target cardinality distributions across the three subsets. (b) Entity category composition for Exact-QG, Gen-QG, and Comp-QG, respectively. (c) Normalized spatial distribution of unique target box centers.

Comp-QG. This subset focuses on quantitative understanding in compositional grounding, where the target is identified together with other entities and their relations. We first select multiple entities in an image and establish their relations according to the visual content. Then, the selected entities, quantities, and relations are then composed into a referring expression. Finally, we perform human verification and reject expressions with incorrect or ambiguous entity-relation compositions.

3.2 Metric

To evaluate quantitative grounding, we consider both localization completeness and quantity understanding. We introduce two metrics, Set-F1@0.5 and Count-Acc. Firstly, let G_i denote the number of referred entity groups for the i-th query. Given the predicted box set $\hat{B}_{i,g}$ and the ground-truth box set $B_{i,g}$ for the g-th entity group, we perform one-to-one matching with an IoU threshold of 0.5. Let $m_{i,g}$ denote the number of matched boxes. Set-F1@0.5 is defined as follows,

$$\text{Set-F1@0.5} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{G_i}\sum_{g=1}^{G_i}\frac{2m_{i,g}}{|\hat{B}_{i,g}| + |B_{i,g}|}, \tag{12}$$

where N denotes the number of phrase queries. It requires predictions to match both the referred entity group and its corresponding instances, thereby penalizing missing and redundant predictions. Subsequently, let $\mathcal{B}_i = \bigcup_{g=1}^{G_i} B_{i,g}$ and $\hat{\mathcal{B}}_i$ denote the overall ground-truth and predicted box sets for the i-th query, respectively. We define Count-Acc as follows,

$$\text{Count-Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{c}_i = c_i], \tag{13}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. c_i and $\hat{c}_i$ are corresponding object counts of $\mathcal{B}_i$ and $\hat{\mathcal{B}}_i$.

4 Experiment

4.1 Experimental Setup

We evaluate Q-VG on the proposed QGround-Bench and two standard WSVG benchmarks, Flickr30K Entities (Plummer et al., 2015) and gRefCOCO (Ding et al., 2026). For QGround-Bench, we use Set-F1@0.5 and Count-Acc defined in Sec. 3.2. For the standard benchmarks, we follow their established evaluation protocols and report Bbox-Acc and Point-Acc on Flickr30K Entities, and Pr@0.5 and N-Acc on gRefCOCO. Detailed dataset statistics, evaluation protocols, and metric definitions are provided in Appendix B.

We compare with representative vision-language pretrained models, including Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), and LocateAnything (Wang et al., 2026b), as well as

Table 2: Comparison on the Exact-QG, Gen-QG, and Comp-QG subsets of QGround-Bench.

	Exact-QG		Gen-QG		Comp-QG	
	Set-F1@0.5	Count-Acc	Set-F1@0.5	Count-Acc	Set-F1@0.5	Count-Acc
V-L Pretrained						
Qwen2.5-VL-7B	46.50	46.20	44.40	40.30	39.60	6.40
Qwen3-VL-4B	61.30	72.60	58.10	57.60	47.90	11.50
Qwen3-VL-8B	61.20	75.30	56.90	59.10	36.00	4.70
LocateAnything-3B	70.00	78.70	62.70	57.60	46.30	2.60
Weakly Supervised						
WWbL-g	6.00	7.20	6.00	7.80	7.80	5.42
VPT	4.00	22.10	4.30	25.10	3.40	3.37
TAS	7.90	8.10	6.80	7.10	7.50	6.20
DAPO	9.60	13.60	9.00	12.10	6.30	7.11
FEU-NET	10.30	19.60	10.50	21.60	6.52	8.60
Q-VG(Ours)	**54.60**	**98.90**	**34.10**	**69.50**	**25.90**	**22.00**

weakly supervised grounding methods including WWbL-g (Shaharabany et al., 2022), VPT (Lin et al., 2024b), TAS (Lin et al., 2024a), DAPO (Lin et al., 2026a), FEU-NET (Lin et al., 2026b), CPL++ (Liu et al., 2026), GKDN (Ke et al., 2025), and LIHE (Shi et al., 2025).

4.2 Implementation Details

We use frozen YOLOv3 DarkNet (Redmon & Farhadi, 2018) and DINOv3 ViT-B/16 (Siméoni et al., 2025) as visual encoders, with the former providing anchor priors. We use GloVe (Pennington et al., 2014) to obtain textual embeddings. Input images are resized to 416×416. Further training and hyperparameter settings are provided in Appendix C.

4.3 Quantitative Results

Table 2 presents the results on the Exact-QG, Gen-QG, and Comp-QG subsets of QGround-Bench. WSVG methods here are trained on the Flickr30K Entities training split. Q-VG consistently achieves the best performance among WSVG methods on both Set-F1@0.5 and Count-Acc across all three subsets. Its Count-Acc is higher than all evaluated V-L pretrained baselines across the three subsets. Furthermore, we evaluate our Q-VG and WSVG baselines on classical datasets, i.e. Flickr30K Entities and gRefCOCO. Table 3 reports the experimental results. The results demonstrate that our network achieves competitive performance under the weakly supervised setting. Q-VG achieves 66.18% Bbox-Acc and 87.03% Point-Acc on the standard Flickr30K evaluation. This result benefits from the synergy between the anchor candidates' boundary localization and the network's quantity reasoning. On gRefCOCO, Q-VG narrows the performance gap with V-L pretrained baselines and achieves competitive N-Acc, including the best result on testB among the V-L pretrained baselines. This is because no-target samples enhance the RECO module's perception of target quantities.

4.4 Ablations Studies

In this section, we investigate how the performance of our network is affected by different model settings. All ablation experiments are based on our Q-VG.

Model Components. We evaluate the RECO module and the two anchor-scoring terms in Eq. 5, namely semantic similarity and the slot prior. As shown in Table 4, adding RECO improves Bbox-Acc and Pr@0.5, since it estimates the target count and constrains the number of selected anchors. Combining semantic similarity with the slot prior achieves the best results on both datasets, showing that semantic alignment and object-level spatial evidence are complementary for anchor selection.

RECO Module. In this section, we analyze the effectiveness of count-aware anchor selection. To examine the effectiveness of our RECO module, we utilize it to predict the amount of target objects. Besides, we adopt a heatmap-generated WSVG method (Lin et al., 2026a) as the baseline. Figure 6 shows that the predicted target counts from our RECO module align more closely with the ground

Table 3: Comparison on Flickr30K Entities and gRefCOCO.

	Flickr30K Entities		gRefCOCO val		gRefCOCO testA		gRefCOCO testB	
	Bbox-Acc	Point-Acc	Pr@0.5	N-Acc	Pr@0.5	N-Acc	Pr@0.5	N-Acc
V-L Pretrained								
Qwen2.5-VL-7B	63.80	66.90	38.20	6.40	47.20	3.40	38.30	4.20
Qwen3-VL-4B	72.93	88.90	72.40	77.20	68.00	70.20	62.20	70.20
Qwen3-VL-8B	74.86	85.90	57.10	67.40	63.10	84.10	60.00	70.10
LocateAnything-3B	72.60	83.20	62.30	42.70	71.50	74.90	69.80	60.80
Weakly Supervised								
WWbL-g	36.35	75.63	1.50	–	1.20	–	2.50	–
VPT	45.60	81.15	3.50	–	4.00	–	5.00	–
TAS	45.63	77.85	1.80	–	1.50	–	2.20	–
DAPO	48.57	85.43	2.50	–	2.80	–	4.00	–
FEU-NET	48.77	84.35	12.30	55.10	18.60	62.10	14.00	52.90
CPL++	–	83.40	–	–	–	–	–	–
GKDN	–	–	32.70	31.50	**35.30**	23.40	27.20	19.80
LIHE	–	–	39.60	67.50	32.70	79.60	35.80	67.10
Q-VG (Ours)	**66.18**	**87.03**	**41.20**	**73.10**	30.80	**80.20**	**38.10**	**73.00**

Table 4: Analysis of Q-VG with different components.

RECO	Anchor Scoring		Flickr30K Entities		gRefCOCO (val)	
	Semantic	Slot prior	Bbox-Acc	Point-Acc	Pr@0.5	N-Acc
✗	✓	✗	42.50	72.84	4.50	–
✓	✓	✗	60.12	82.56	33.20	72.50
✓	✗	✓	58.00	81.99	35.40	72.60
✓	✓	✓	**66.18**	**87.03**	**41.20**	**73.10**

Table 5: Analysis of Q-VG with different losses.

Loss Terms			Flickr30K Entities		gRefCOCO (val)	
L_{mcl}	L_{rec}	L_{cnt}	Bbox-Acc	Point-Acc	Pr@0.5	N-Acc
✓	✗	✗	22.28	42.74	5.40	0.10
✓	✓	✗	25.47	46.66	6.60	0.10
✓	✗	✓	64.75	85.61	40.10	71.90
✓	✓	✓	**66.18**	**87.03**	**41.20**	**73.10**

truth. It indicates that the RECO module could accurately predict the amount of the target objects. In contrast, the proportion of correctly predicted objects by DAPO is relatively small, as the number of bounding boxes inferred from the heatmap tends to be unstable.

Loss. Table 5 evaluates the contributions of the three training objectives. L_{mcl} provides set-level phrase–region alignment by increasing the aggregate probability mass of selected positive anchors. Adding L_{rec} to L_{mcl} yields a modest improvement, whereas introducing L_{cnt} substantially improves both grounding and no-target recognition. This comparison underscores the importance of explicit cardinality supervision for quantity-aware grounding.

4.5 Qualitative Results

Figure 7 shows qualitative comparisons with representative baselines. In (a), aided by the RECO module, Q-VG predicts target counts consistent with the annotations and grounds the entities mentioned in the phrase, such as the drinks and couch. In (b), Q-VG grounds phrase-relevant entities beyond subjects, such as the frisbee, and is able to further recognize referred objects that are misidentified by V-L pretrained models. These results demonstrate Q-VG's ability to jointly reason about target quantities and ground multiple referred objects.

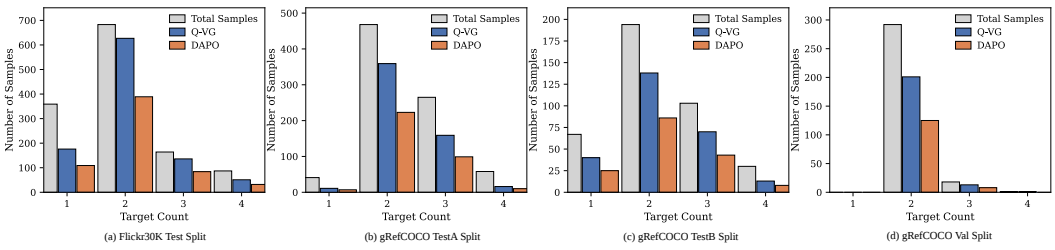


Figure 6: Count-prediction for Q-VG and DAPO on Flickr30K Entities and gRefCOCO.

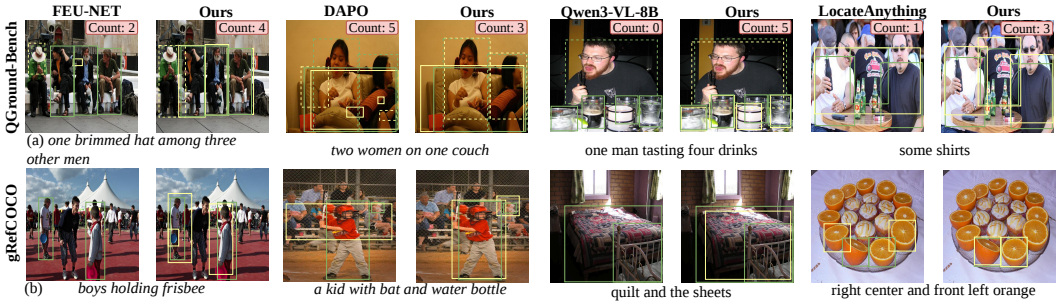


Figure 7: Visualizations of Q-VG and common visual grounding baselines. The yellow and green boxes are the predicted and ground truth ones, respectively.

5 Related Works

5.1 Weakly Supervised Visual Grounding

WSVG localizes objects from image-text pairs without bounding-box annotations. Existing methods are broadly divided into anchor-free and anchor-based approaches. Anchor-free methods predict localization maps from cross-modal features (Zhang et al., 2018; Akbari et al., 2019; Arbelle et al., 2021; Lin et al., 2024b; He et al., 2025; Lin et al., 2026a). Anchor-based methods generate object proposals with pretrained detectors and learn phrase-region matching. Datta et al. (2019) consider cross-modal phrase-region alignment, while Liu et al. (2021) incorporate contextual relations among proposals. Recent methods, including Jin et al. (2023), Luo et al. (2024), and Chen et al. (2025) further improve representation learning, pseudo-label generation, and query-object matching. However, previous methods mainly identify relevant regions and overlook target quantity. This often causes a mismatch between the predicted boxes and the number of referred entities.

5.2 Compositional Grounding Benchmarks

Grounding benchmarks have evolved to evaluate increasingly complex linguistic compositions and referent structures. Flickr30K Entities (Plummer et al., 2015) provides phrase-level annotations for natural image descriptions, where one phrase corresponds to one or multiple objects. Cops-Ref (Chen et al., 2020) and Ref-Reasoning (Yang et al., 2020) introduce complex referring expressions to evaluate compositional grounding with attributes and object relations. gRefCOCO (Ding et al., 2026) extends conventional single-target grounding to zero-, single-, and multi-target expressions. ARPGrounding (Zeng et al., 2024) and FineCops-Ref (Liu et al., 2024) provide fine-grained evaluations of compositional reasoning, including attributes, relations, priorities, and multi-hop dependencies. However, existing benchmarks rarely isolate quantity as a compositional cue for identifying the complete set of referents.

6 Conclusion and Future Work

In this paper, we propose Q-VG, a count-aware model for WSVG. Specifically, we first introduce a count-aware anchor selection scheme. This scheme leverages the RECO module to estimate the count of target objects, ensuring that the number of predicted instances aligns with the query phrase. We also design an anchor-text matching objective, aligning multiple positive objects with the phrase. Furthermore, we introduce QGround-Bench, a quantitative WSVG benchmark with three subsets. Finally, we conduct experiments on QGround-Bench, Flickr30K Entities, and gRefCOCO datasets. The experimental results show that our Q-VG achieves competitive performance, which demonstrates quantitative reasoning and grounding in multi-target scenarios.

In the future, we will investigate Q-VG in more realistic settings, such as crowded scenes, ambiguous referring expressions, and unseen object categories. We also plan to extend Q-VG to interactive scenarios such as embodied perception, where models need to identify both the referred objects and their quantities for reliable actions.

REFERENCES

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. Multi-level multimodal common semantic space for image-phrase grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12476–12486, 2019.

Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1801–1812, 2021.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Xiaofu Chen, Yaxin Luo, Gen Luo, Jiayi Ji, Henghui Ding, and Yiyi Zhou. Dvin: Dynamic visual routing network for weakly supervised referring expression comprehension. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14347–14357, 2025.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10083–10092. IEEE, 2020.

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2601–2610, 2019.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Yu-Gang Jiang. Grex: Generalized referring expression segmentation, comprehension, and generation. *International Journal of Computer Vision*, 134(2):79, 2026.

Ruozhen He, Ziyan Yang, Paola Cascante-Bonilla, Alexander C. Berg, and Vicente Ordonez. Learning from synthetic data for visual grounding. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025. URL https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_22/paper.pdf.

Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2681–2690, 2023.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1780–1790, 2021.

Jingcheng Ke, Jia Wang, Xu Lu, Fangyuan Lei, YuYang, Zhenggen Li, and Jie Wen. Wgrec: Weakly supervised generalized referring expression comprehension empowered by large language model. 2025. URL https://openreview.net/forum?id=URMih4CWyR.

Dongdong Kuang, Richong Zhang, Zhijie Nie, Junfan Chen, and Jaein Kim. Momentum pseudo-labeling for weakly supervised phrase grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24348–24356, 2025.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8778–8786, 2025.

Pengyue Lin, Ruifan Li, Yuzhe Ji, Zhihan Yu, Fangxiang Feng, Zhanyu Ma, and Xiaojie Wang. Triple alignment strategies for zero-shot phrase grounding under weak supervision. In *ACM Multimedia 2024*, 2024a.

Pengyue Lin, Zhihan Yu, Mingcong Lu, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. Visual prompt tuning for weakly supervised phrase grounding. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7895–7899, 2024b.

Pengyue Lin, Yanyang Hu, Xinjing Liu, Wenqi Jia, Fangxiang Feng, and Ruifan Li. Diffusion-assisted progressive learning for weakly supervised phrase localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32024–32032, 2026a.

Pengyue Lin, Ruifan Li, Fangxiang Feng, Lun Ke, Zhanyu Ma, and Xiaojie Wang. A fine-grained entity understanding network for weakly supervised phrase grounding. *Pattern Recognition*, 180: 114075, 2026b.

Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15440–15457, 2024.

Yang Liu, Jiahua Zhang, Yue Wu, Zijing Zhao, Qingchao Chen, and Yuxin Peng. Confidence-aware pseudo-label self-correction for weakly supervised visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5612–5621, 2021.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Yaxin Luo, Jiayi Ji, Xiaofu Chen, Yuxin Zhang, Tianhe Ren, and Gen Luo. Apl: Anchor-based prompt learning for one-stage weakly supervised referring expression comprehension. In *European Conference on Computer Vision*, pp. 198–215. Springer, 2024.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

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

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv:1804.02767*, 2018.

Tal Shaharabany, Yoad Tewel, and Lior Wolf. What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. *Advances in Neural Information Processing Systems*, 35:28222–28237, 2022.

Xianglong Shi, Silin Cheng, Sirui Zhao, Yunhan Jiang, Enhong Chen, Yang Liu, and Sebastien Ourselin. Lihe: Linguistic instance-split hyperbolic-euclidean framework for generalized weakly-supervised referring expression comprehension. *arXiv:2511.12020*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv:2508.10104*, 2025.

Fei Tang, Zhangxuan Gu, Zhengxi Lu, Xuyang Liu, Shuheng Shen, Changhua Meng, Wen Wang, Wenqi Zhang, Yongliang Shen, Weiming Lu, et al. Gui-g^2: Gaussian reward modeling for gui grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33214–33222, 2026.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22274–22284, 2025.

Jiacong Wang, Zijian Kang, Haochen Wang, Xiao Liang, Ya Wang, Jiawen Li, Bohong Wu, Jiao Ran, Haiyong Jiang, Chao Feng, et al. Vgr: Visual grounded reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 89506–89528, 2026a.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 336–357, Cham, 2026b. Springer Nature Switzerland. ISBN 978-3-032-37718-0.

Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9949–9958. IEEE, 2020.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14141–14151, 2024.

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126 (10):1084–1102, 2018.

APPENDIX

The supplementary material provides additional experimental details and qualitative results to support the findings of the main paper. Specifically, it includes:

- **AI Disclosure**: The statement for using AI tools.
- **Experimental Details**: Statistics for the standard WSVG datasets and QGround-Bench.
- **Implementation Details**: Our network's training details.
- **Ablation Study**: The ablation experiments for ablating different hyperparameters.
- **Visualization Analysis and Failure Cases** : More quantitative results and failure cases for our Q-VG network.
- **Computational Cost**: The computational statistics for our network and baselines.

A AI DISCLOSURE

In this work, we used generative AI tools to polish the text and filter the benchmark dataset. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

B EXPERIMENTAL DETAILS

B.1 DATASETS AND EVALUATION PROTOCOLS

Flickr30K Entities (Plummer et al., 2015) contains 456K phrases describing bounding boxes in 31K images. Each image includes five captions, and each caption contain several phrases. We follow the protocol in MG (Akbari et al., 2019) and select 1000 images from the test split to evaluate.

gRefCOCO (Ding et al., 2026) contains 258K expressions describing bounding boxes in 19,994 images. Each expression targets single, multiple, or no objects.

QGround-Bench is constructed on the Flickr30K Entities test split to evaluate quantitative grounding under different forms of quantity expressions and entity compositions. It consists of three subsets, namely Exact-QG, Gen-QG, and Comp-QG. Detailed statistics are reported in Table 6.

1) Annotation scope. All entities explicitly referred to by a query are included in the grounding annotations. The terms *target* and *context* indicate their semantic roles in the expression. Exact-QG and Gen-QG contain a single referred entity group, whereas Comp-QG contains multiple entity groups connected by semantic relations. For example, "one man tasting four drinks" contains two entity groups, and the annotation includes one man and four drinks, resulting in five referred boxes in total.

2) Exact-QG. We first select Flickr30K Entities phrases associated with multiple bounding boxes and obtain their cardinalities directly from the original annotations. The phrases are then rewritten with explicit cardinal numbers corresponding to the annotated object counts. Since additional objects of the same category may appear in an image without being referred to by the caption, we remove samples for which the rewritten expression does not uniquely identify the annotated object set. The remaining samples are verified for consistency between the rewritten quantity, phrase semantics, and original bounding boxes.

3) Gen-QG. Starting from Exact-QG, we replace explicit cardinal expressions with generalized or lexicalized quantity expressions while preserving the original referred object set. The transformations include expressions such as "some", "several", "multiple", "both", and "a pair of", subject to their cardinality compatibility. An MLLM is first used to screen whether the transformed query remains semantically compatible with the original referents. Human annotators then verify whether the generalized expression can still be reasonably associated with the same annotated object set. Samples introducing ambiguous or inconsistent referents are discarded.

4) Comp-QG. Comp-QG is constructed by composing multiple annotated entity groups with semantic relations. For each candidate pair, we first extract the relation from the original caption whenever

Table 6: Statistics of the three subsets in QGround-Bench.

Subset	Images	Queries	Target Boxes	Avg. Targets	Max Targets	Avg. Entities	Avg. Query Length
Exact-QG	628	1,553	5,055	3.255	10	1.000	3.672
Gen-QG	623	6,237	18,083	2.899	10	1.000	5.227
Comp-QG	596	4,425	20,781	4.696	16	2.000	7.064

possible. Otherwise, an MLLM is used to infer a visually supported relation. Quantitative expressions are then assigned to the constituent entity groups, including both exact and generalized forms. We retain the original bounding boxes for each entity group separately, while their union defines the complete set of referred objects. Note that the presence of additional same-category instances does not directly invalidate a sample. It is retained when the full compositional expression, together with its relation, sufficiently identifies the annotated referents. Expressions with unsupported relations or ambiguous entity–relation compositions are removed.

5) Human verification protocol. Automatic and MLLM-assisted procedures are used only for candidate generation and preliminary ambiguity screening. Each selected audit item is independently reviewed by two annotators (A1 and A2), who assign a keep/reject decision together with subset-specific checks and rejection codes. For Exact-QG, they verify quantity consistency and referent uniqueness; for Gen-QG, they additionally assess compatibility between the generalized expression and the annotated object set; and for Comp-QG, they verify the constituent entity groups, quantities, and visual relation. Disagreements are forwarded to a separate adjudication stage, which determines the final label.

The two independent decisions yield a raw agreement of 91.1% before adjudication; this statistic is computed from the original A1/A2 votes rather than inferred from the final acceptance field. After adjudication, 12,215 samples were retained and 1,062 were rejected, corresponding to an acceptance rate of 92.0%. The rejected samples mainly involved ambiguous referents 488, inconsistent quantitative expressions 361, or unsupported/ambiguous relations 213.

B.2 Evaluation Metrics

Flickr30K Entities. We use the following metrics, including "pointing game" accuracy (i.e., Point-Acc) (Liu et al., 2021) and bounding box accuracy (i.e., Bbox-Acc) (Kamath et al., 2021). The former measures the percentage of the center point of the predicted box that lies within the bounding box ground truth. The latter measures the percentage of bounding boxes that have an IoU greater than 0.5 for the testing set of "image-phrase" pairs.

gRefCOCO. We utilize Pr@0.5 and N-Acc to evaluate baselines. The former computes the percentage of samples that have the F_1 score of 1 with the IoU threshold set to 0.5. The latter assesses the model's proficiency in no target identification as $\frac{TP}{TP+FN}$. Here, for a no-target sample, prediction without any bounding box is considered a true positive (TP), otherwise, it is a false negative (FN).

QGround-Bench. Set-F1@0.5 evaluates localization within the attributed entity groups, while Count-Acc evaluates the cardinality of the full prediction set. Specifically, Set-F1@0.5 measures whether predicted boxes are assigned to the correct entity groups and match the corresponding instances, whereas Count-Acc only compares the number of predicted and ground-truth boxes. Since these baselines originally produce localization heatmaps rather than an explicit set of bounding boxes, we apply a shared multi-box decoding procedure to obtain predictions for Set-F1@0.5 and Count-Acc. Specifically, each heatmap is thresholded at 0.5 of its maximum response, and connected high-response regions are converted into bounding boxes by contour extraction. The generated boxes are further filtered using the same relative confidence threshold and NMS setting (IoU $= 0.05$) for all heatmap-based baselines, which is a fixed shared setting. Because heatmap-based methods require an additional connected-component conversion to produce an explicit box set, their results can remain sensitive to this decoding interface. We therefore interpret Table 2 as a comparison under a standardized multi-box protocol rather than a strictly architecture-controlled comparison.

Table 7: Sensitivity of heatmap-based baselines to the NMS IoU threshold on QGround-Bench. Each entry reports Set-F1@0.5 / Count-Acc.

Method	NMS	Exact-QG	Gen-QG	Comp-QG
DAPO	0.05	**9.60 / 13.60**	**9.00 / 12.10**	**6.30 / 7.11**
DAPO	0.10	9.70 / 12.80	9.10 / 11.50	6.40 / 6.70
DAPO	0.30	10.30 / 10.10	9.60 / 9.80	7.00 / 5.90
DAPO	0.50	10.90 / 8.20	10.00 / 8.50	7.50 / 5.10
FEU-NET	0.05	**10.30 / 19.60**	**10.50 / 21.60**	**6.52 / 8.60**
FEU-NET	0.10	10.50 / 19.10	10.70 / 21.00	6.70 / 8.40
FEU-NET	0.30	11.40 / 17.20	11.30 / 19.10	7.40 / 7.90
FEU-NET	0.50	12.20 / 15.00	12.00 / 17.50	8.10 / 7.20

Table 8: Quantitative supervision used for $\mathcal{L}_{\mathrm{cnt}}$.

Type	**Expression / Example**	$\mathbf{C}_{\mathrm{low}}$	$\mathbf{C}_{\mathrm{high}}$	**I**
Exact	one . . . twenty / number k	k	k	1
Dual	both, pair, couple	2	2	1
Vague-small	a few, several	2	–	0
Vague	some, multiple	2	–	0
Vague-large	many, lots, various	3	–	0
Lower	at least k	k	–	0
Strict-lower	more than k / over k	$k+1$	–	0
Upper	at most k / no more than k	1	k	1
Range	between a and b	$\min(a,b)$	$\max(a,b)$	1
No-target	annotated empty referent set	0	0	1
Comp. exact	one man holding four drinks	$1+4=5$	$1+4=5$	1
Comp. vague	some people near two cars	$2+2=4$	–	0
Comp. vague	several dogs and a few cats	$2+2=4$	–	0

For each Comp-QG query q with entity-group phrases $\{P_g\}_{g=1}^{G}$, every method predicts B_{full} from q and B_g from each P_g. We enumerate all pairs between B_{full} and B_g with IoU at least 0.5, sort them in descending IoU order, and greedily perform one-to-one matching. Each full-query box and each group-query box can be matched at most once. A matched full-query box is assigned to the corresponding group, while unmatched full-query boxes are excluded only from group-wise localization scoring. This produces $\hat{B}_g$ for Set-F1@0.5. Count-Acc instead compares $|B_{\mathrm{full}}|$ with the cardinality of the ground-truth box union. No ground-truth box or count is used to generate predictions or perform group attribution. The same protocol is applied to Q-VG, heatmap-based WSVG baselines, and V-L pretrained models.

Table 7 varies only the NMS threshold while keeping the checkpoints, heatmap threshold, contour extraction, confidence filtering, and evaluation samples fixed. Relaxing NMS from 0.05 to 0.50 moderately improves Set-F1 but reduces Count-Acc because more components are retained. The best sensitivity result remains substantially below Q-VG on every subset, indicating that aggressive suppression is not the primary cause of the performance gap. We therefore retain the fixed 0.05 protocol in Table 2 rather than selecting a threshold on the test set.

C Implementation Details

We resize the input image to 416×416. The maximum length of the input phrase is set to 15 for Flickr30K Entities. For visual encoders, we utilize DINOv3 ViT-B/16 (Siméoni et al., 2025) and DarkNet (Redmon & Farhadi, 2018) from YOLOv3. The latter provides explicit anchor selection prior for bounding box output. For the text encoder, we initialize a 300-dimensional embedding layer with GloVe vectors from spaCy's `en_vectors_web_lg` and keep it frozen. A randomly initialized 512-dimensional GRU and the subsequent self-attention pooling layers are optimized during weakly supervised training. For detection head, we utilize the pretrained YOLO head. The visual encoders, YOLO head, and GloVe embedding layer are frozen, whereas the GRU, self-attention pooling layers,

and remaining task-specific modules are trainable. During weakly supervised training, we use the Adam optimizer with a learning rate of 1e-4. The batch size is set to 128. The training process lasts for 25 epochs. For loss weights, we set λ_r and λ_c as 0.1. The temperature parameter τ is set as 0.07. For the RECO module, we set the slot capacity K to 16 and perform $T = 3$ iterative slot updates.

Scalar Slot Prior. The slot attention in Eq. (3) contains K channels and therefore cannot be added directly to the scalar anchor–phrase similarity. We first aggregate the channels into a count-weighted scalar saliency map:

$$\bar{A}(u, v) = \sum_{k=1}^{K} \tilde{p}_k A^{\text{slot}}_{u,v,k}, \qquad \tilde{p}_k = \frac{p_k}{\sum_{j=1}^{K} p_j + \epsilon}. \tag{14}$$

The slot-prior term in Eq. 5 is then obtained by spatially averaging $\bar{A}$ inside the region Ω_a of each candidate anchor. This aggregation preserves the spatial evidence of slots predicted as target objects while producing a scalar score for anchor ranking.

We also reshape the DINO patch tokens into a 2-D feature map, bilinearly resize it to the spatial resolution of the stride-8 YOLO feature map, and project its channels to D using a 1×1 convolution. For each candidate anchor, we rescale its coordinates from the input image to the same feature grid and average-pool the slot saliency map $\bar{A}$ within the corresponding region.

No-target Supervision. For datasets that provide no-target annotations, an empty referent set overrides surface quantitative words in the expression. We set $v_n = 1$, $C_{n,\text{low}} = C_{n,\text{high}} = 0$, and $\mathbb{I}_n = 1$ in Eq. 9. Consequently, the count branch is explicitly optimized toward $N_n = 0$. Since $\mathcal{A}^\star_n$ is empty, this sample is excluded from the outer average of $\mathcal{L}_{\text{mcl}}$ but can still contribute candidate anchors as cross-image negatives. During inference, $\hat{N} = 0$ returns an empty prediction set.

Count Supervision. Table 8 provides the mapping from quantitative expressions to count constraints in Eq. 9. Expressions without reliable quantitative cues use $v_n = 0$ and do not contribute to $\mathcal{L}_{\text{cnt}}$; all valid expressions and annotated no-target samples use $v_n = 1$. For a single entity group, its quantitative expression is mapped to $(C_{\text{low}}, C_{\text{high}}, \mathbb{I})$. For a compositional query containing G referred entity groups, we aggregate the group-wise constraints as follows:

$$C_{\text{low}} = \sum_{g=1}^{G} c^g_{\text{low}}, \qquad \mathbb{I} = \prod_{g=1}^{G} \mathbb{I}_g, \tag{15}$$

and define

$$C_{\text{high}} = \begin{cases} \sum_{g=1}^{G} c^g_{\text{high}}, & \mathbb{I} = 1, \\ \text{undefined}, & \mathbb{I} = 0. \end{cases} \tag{16}$$

When $\mathbb{I} = 1$, the upper-bound term is activated; when $\mathbb{I} = 0$, only the lower-bound constraint is applied. Thus, the counting objective supervises the cardinality of the union of all referred entity groups rather than only a single entity group.

Count Discretization. RECO outputs a continuous referring-object count as

$$N = \sum_{k=1}^{K} p_k, \tag{17}$$

where p_k denotes the existence probability of the k-th slot. During inference, we discretize the continuous count by

$$\hat{N} = \text{clip}\left(\text{round}(N), 0, \min\left(N_{\max}, |\mathcal{A}_{\text{NMS}}|\right)\right), \tag{18}$$

where $\mathcal{A}_{\text{NMS}}$ denotes the candidate anchors remaining after NMS and $N_{\max} = K$. We then select the Top-$\hat{N}$ anchors according to Eq. (6). When $\hat{N} = 0$, the model returns an empty prediction set.

Training Protocol of WSVG Baselines. Existing anchor-based WSVG methods are primarily designed for conventional single-referent grounding and typically return the highest-scoring proposal. QGround-Bench instead requires variable-cardinality box-set prediction; adding a count predictor or a manually specified Top-k rule would alter their original inference protocol. We therefore focus the QGround-Bench comparison on methods that can be consistently converted to set prediction, and use controlled Q-VG ablations to isolate the contribution of quantitative modeling. We reproduce WWbL-g, VPT, TAS, DAPO, and FEU-NET under the weakly supervised setting. Their visual backbones are initialized from models pretrained on MSCOCO. For Flickr30K Entities and QGround-Bench, all reproduced baselines are trained on the Flickr30K Entities training split and directly evaluated on the corresponding test sets. For gRefCOCO, we train these baselines on its training split following the supervision used by each original method. Methods with a counterfactual branch use the corresponding counterfactual samples and expression-level labels, while the others use only image–text pairs. No region-level annotations from the target grounding dataset are used during weakly supervised training. For CPL++, GKDN, and LIHE, we directly report the results from their original papers. These results are therefore used as reference comparisons rather than strictly controlled comparisons under an identical training protocol.

Evaluation Protocol of V-L Pretrained Models. We evaluate Qwen2.5-VL, Qwen3-VL, and LocateAnything under a native multi-object grounding setting. Each model receives only the RGB image and the referring expression, without GT boxes, target counts, or numeral-based Top-k post-processing. For Qwen2.5-VL and Qwen3-VL, we use a multi-box instruction that requires the model to determine the number of referred objects and return all bounding boxes in JSON format. For Comp-QG, the full-query prompt explicitly instructs the model to ground all explicitly mentioned entity groups, including contextual entities. LocateAnything follows its official multi-instance grounding prompt. We parse all predicted boxes and map their coordinates to the original image size. Invalid or degenerate boxes are removed, while malformed outputs are treated as empty predictions. For Comp-QG, predictions from the full query and the dataset-provided entity-group phrases are evaluated by the shared prediction-only attribution protocol described above. The prompts for Qwen-VL and LocateAnything are shown below, respectively.

►Qwen-VL Prompt for QGround-Bench:
►Input:
You are a visual grounding model. Given an image and a referring expression, localize the complete set of objects referred to by the expression.
Infer from the image and expression how many objects are referred to, and return exactly one bounding box for each referred object without duplicates. For a compositional expression, return boxes for all explicitly mentioned entity groups, including contextual entities involved in relations. Do not output a separate count field.
Referring Expression: {query}
Image Size: width={width}, height={height}
►Output:
Reply with only one JSON object, with no markdown fences or extra text.
Format exactly as:
`{"boxes": [[x1, y1, x2, y2], ...]}`
Coordinates are absolute pixel coordinates on the original image, with the top-left corner as the origin. Each box must satisfy $x_1 < x_2$ and $y_1 < y_2$ and lie within the image boundaries. If no object is referred to, return `{"boxes": []}`.

►LocateAnything Prompt:
Locate all the instances that match the following description: {query}.

D Ablation Study

In this section, we supplement more ablation studies for Q-VG.

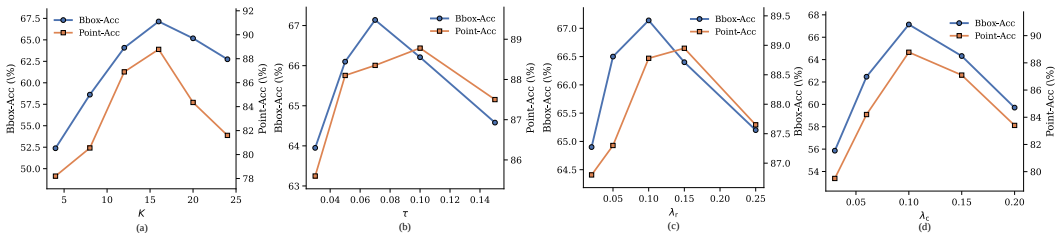


Figure 8: Hyperparameter ablation for Q-VG on the Flickr30K Entities test split. For each panel we vary a single quantity while fixing all other training settings to their defaults, and report Bbox-Acc (left axis, blue) and Point-Acc (right axis, orange). (a) Maximum slot count K in the RECO module. (b) Temperature τ in the MCL loss. (c) Reconstruction loss weight λ_{r}. (d) Counting loss weight λ_{c}
.

Component Settings. For the variants with RECO in Table 4, the predicted count determines the number of selected anchors. The Semantic and Slot prior columns specify the terms used for anchor ranking in Eq. 5, while L_{mcl} remains active for both single-score variants. Removing RECO also removes the slot prior, L_{rec}, and L_{cnt}. This variant returns only the highest-scoring anchor and cannot produce an empty set; therefore, its N-Acc is not reported.

► Qwen-VL Prompt for Flickr30K Entities and gRefCOCO:
►Input:
You are a visual grounding model. Given an image and a referring expression, localize only the object(s) referred to by the expression.
Do not return entities that are mentioned only as context, attributes, or relational cues. For example, for "the man next to two dogs", return the box(es) of the man only, not the dogs.
If the expression refers to multiple instances, return one bounding box for each instance. Do not output a separate count field and do not use external candidate boxes.

Referring Expression: {query}
Image Size: width={width}, height={height}
►Output:

Reply with only one JSON object, with no markdown fences or extra text.
Format exactly as:
`{"boxes": [[x1, y1, x2, y2], ...]}`
Coordinates are absolute pixel coordinates on the original image, with the top-left corner as the origin. Each box must satisfy $x_1 < x_2$ and $y_1 < y_2$ and lie within the image boundaries.
If no object is referred to, return:
`{"boxes": []}`.

Hyperparameters. We ablate the hyperparameters K for the RECO module, and $\tau, \lambda_r, \lambda_c$ for loss functions. We analyze the sensitivity of Q-VG to key hyperparameters in our weakly supervised grounding network. All experiments follow the Flickr30K Entities test split, using Bbox-Acc and Point-Acc as the two metrics for localization quality. Figure 8 summarizes this analysis. Specifically, we ablate the maximum slot capacity K in the phrase-conditioned slot attention module, the temperature τ used in L_{mcl}, the reconstruction weight λ_{r}, and the counting weight λ_{c}. Note that the slot capacity K strongly affects model behavior. If K is low, some referred instances may never be accounted for. If K is high, the model incurs unnecessary computational overhead and becomes more susceptible to mistaken judgments.

RECO Module. To investigate whether RECO relies on both visual and textual information for quantitative reasoning, we conduct modality ablations on Gen-QG. While Exact-QG mainly evaluates explicit cardinality following, Gen-QG provides a more direct diagnostic of visually grounded count inference. We compare the full RECO with three controlled variants. Text-only removes visual feature aggregation and predicts the target count only from the referring expression. Vision-only removes phrase conditioning and estimates the count only from DINO visual features. Global

Table 9: Ablation of modality information for quantitative grounding.

Method	Gen-QG		Gen-Vague	
	Count-Acc ↑	Set-F1 ↑	Target-Count-Match ↑	Set-F1 ↑
Text-only Count	67.3	32.6	48.7	29.4
Vision-only Count	28.5	24.2	21.6	22.3
Global V-L Count	68.2	32.9	53.4	30.8
Full RECO	**69.5**	**34.1**	**58.6**	**32.6**

V-L Count replaces the slot-based count readout with a global vision-language counting head:

$$N_{\text{global}} = K \cdot \sigma(\text{MLP}\left[\text{GAP}(f_d); f_p\right]). \tag{19}$$

For this variant, the phrase-conditioned slots are retained only to compute the same slot prior used for anchor ranking; the count supplied to Top-$\hat{N}$ selection is predicted by Eq. 19. Thus, the candidate anchors, training objectives, joint anchor scoring, and decoding procedure remain unchanged. We further evaluate these variants on Gen-Vague, a 1,439-query subset of Gen-QG with `family=vague`. It is rewritten from Exact-QG seeds using the deterministic mapping $\geq 5 \rightarrow some/several/many$. Expressions with deterministic lexical counts are excluded. Because these expressions do not reveal an exact count from text alone, we refer to exact cardinality agreement on this subset as Target-Count-Match.

As shown in Table 9, Text-only achieves comparable performance on Gen-QG, since some generalized expressions still contain strong quantitative cues. However, Text-only obtains only 48.7% Target-Count-Match on Gen-Vague, 18.6 points below its 67.3% Count-Acc on Gen-QG. In contrast, Full RECO achieves 58.6% Target-Count-Match and 32.6% Set-F1 on Gen-Vague, outperforming Text-only by 9.9 and 3.2 points, respectively. Vision-only performs substantially worse because visual information alone cannot determine which entities are referred to by the query. These results show that RECO combines textual quantity cues with visual object information, particularly for vague quantitative expressions.

DINOv3 provides the dense visual representation used by RECO and is therefore treated as part of the proposed architecture rather than an independently removable component. To control for backbone capacity, all Q-VG ablations retain the same frozen visual encoders. We further separate the effect of counting supervision from Top-$\hat{N}$ decoding. As shown in Table 10, using the GT count improves the model without L_{cnt} by 7.5, 6.5, and 6.8 points on Exact-QG, Gen-QG, and Comp-QG, respectively, showing the importance of output cardinality. Moreover, under the same GT-count decoding, Full Q-VG still improves Set-F1 by 2.7, 4.7, and 4.7 points. This indicates that L_{cnt} not only provides a more reliable count for Top-$\hat{N}$ decoding, but also benefits count-aware positive anchor selection during training.

To determine whether the gains of RECO arise merely from introducing a count predictor, we compare it with the global vision-language counting head under the same downstream grounding pipeline. Global V-L Count performs comparably to RECO on Gen-QG, where many expressions contain informative linguistic quantity cues. However, its Target-Count-Match decreases to 53.4% on Gen-Vague, compared with 58.6% for Full RECO. Its corresponding Set-F1 is 30.8%, compared with 32.6% for Full RECO. This result suggests that global visual-language representations provide useful counting information, whereas phrase-conditioned object slots are more effective when the exact cardinality must be inferred from individual visual instances.

Table 10: Effect of counting supervision and decoding count on QGround-Bench. We report Set-F1@0.5 on all subsets.

Training	Decoding Count	Exact-QG	Gen-QG	Comp-QG
w/o $\mathcal{L}_{\text{cnt}}$	Pred. Count	44.5	25.0	15.2
w/o $\mathcal{L}_{\text{cnt}}$	GT Count	52.0	31.5	22.0
Global V-L Count	Pred. Count	53.7	32.9	23.4
Full Q-VG	Pred. Count	54.6	34.1	25.9
Full Q-VG	**GT Count**	**54.7**	**36.2**	**26.7**

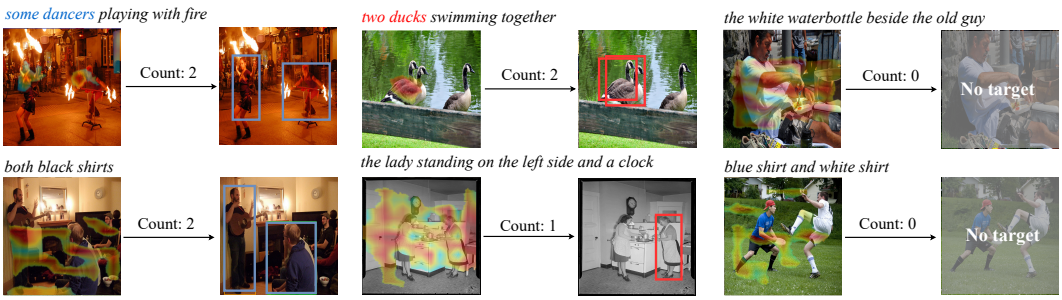


Figure 9: Illustration of counting prediction by our RECO module to guide phrase grounding on Flickr30K Entities and gRefCOCO. The examples in first row shows the correct predictions, while the second row shows the failure cases.

Under predicted-count decoding, Global V-L Count obtains Set-F1 scores of 53.7%, 32.9%, and 23.4% on Exact-QG, Gen-QG, and Comp-QG, respectively. Full RECO improves these results to 54.6%, 34.1%, and 25.9%. The larger improvement on Comp-QG indicates that object-level counting is particularly useful when multiple entity groups must be resolved jointly.

Relation Intervention. To examine whether Q-VG uses compositional relations rather than merely aggregating independently grounded entity groups, we perform an inference-only intervention on Comp-QG. The entity categories and quantitative expressions are preserved in all settings. Relation removed deletes the relational phrase and connects the entity groups with a conjunction, whereas Relation shuffled replaces the original relation with an incompatible relation sampled from another query. No model parameter is updated for this analysis.

Table 11: Effect of relation intervention on Comp-QG.

Query Setting	Set-F1@0.5	Count-Acc
Relation shuffled	18.73	20.91
Relation removed	21.84	21.62
Full relation	25.90	22.00

As shown in Table 11, removing the relation reduces Set-F1 from 25.90% to 21.84%, and replacing it with an incompatible relation further decreases Set-F1 to 18.73%. In contrast, Count-Acc changes only marginally because the entity categories and quantity expressions remain unchanged. This asymmetric degradation indicates that relational information primarily affects the selection of the correct set members rather than the predicted total cardinality.

E Visualization Analysis and Failure Cases

In this section, we present the visualization of counting results generated by our RECO module on benchmark datasets. As shown in the top row of Figure 9, our RECO module is capable of providing precise quantitative supervision for the grounding network. For instance, when the RECO module correctly focuses on the candidate regions ("*some dancers playing with fire*", "*two ducks swimming together*"), it can determine the final number of bounding boxes and position. However, when the RECO module predicts zero, the network receives no boxes. Since the network is optimized for non-target supervision, it can thus be improved with the feedback signals provided by the RECO module. The bottom row presents typical failure cases. In the first example, the module counts the target "*both black shirts*" correctly, and the network grounds incorrect regions. This is because the network fails to capture the semantic nuances while aligning with the phrase query. In the latter two examples, the RECO module receives insufficient supervision even though it counts the wrong targets. This is attributed to the nature of weak supervision, which may fail to precisely constrain the target number based on phrase queries in complex scenarios. Additional qualitative results are shown in Figures 10–13.

Table 12: Comparison of training and inference cost. Inference time and IPS are measured on a single NVIDIA A6000 GPU. "–" denotes unavailable or non-comparable statistics. IPS: images per GPU second.

Method	FLOPs	Trainable Params	Inference Time	IPS
V-L Pretrained				
Qwen2.5-VL-7B	–	–	2000 ms	0.50
Qwen3-VL-4B	–	–	7350 ms	0.14
Qwen3-VL-8B	–	–	9050 ms	0.11
LocateAnything-3B	–	–	900 ms	1.11
Weakly Supervised				
WWbL-g	113.35 G	17.08 M	5.97 ms	167.54
TAS	119.77 G	17.08 M	6.05 ms	165.29
VPT	158.47 G	17.08 M	44.16 ms	22.64
Q-VG	151.24 G	23.06 M	13.62 ms	73.42

F Computational Cost

Table 12 compares the computational cost of Q-VG with several baselines. Inference latency and IPS are measured on a single NVIDIA A6000 GPU. For the WSVG models, FLOPs are reported for an end-to-end 416×416 forward pass, including frozen visual encoders. Parameter counts in the table denote trainable parameters; for Q-VG, the complete checkpoint contains approximately 211.5M parameters, while 23.06M task-side parameters are trainable. As shown in the table, Q-VG incurs higher FLOPs and longer inference latency than lighter pipelines such as WWbL-g (e.g., $5.97\,\mathrm{ms} \rightarrow 13.62\,\mathrm{ms}$ per image, and IPS $167.54 \rightarrow 73.42$ in our profiling setup). The overhead mainly comes from an additional frozen DINOv3 ViT-B/16 forward pass for dense visual features, together with the lightweight counting and slot-attention branch and fused anchor scoring used during training and inference.

Nevertheless, we view this trade-off as acceptable. The extra computation yields consistent gains in weakly supervised grounding quality and exhibits more stable behavior under counting and no-target aware decoding (Table 3). For challenging weakly supervised visual grounding with counting-aware predictions, we believe this accuracy and efficiency trade-off is justified.



Figure 10: Qualitative results of Q-VG on Exact-QG, demonstrating quantitative grounding with explicit cardinalities. The red and green boxes are the predicted and ground truth ones, respectively. Count denotes the predicted number of referred objects.

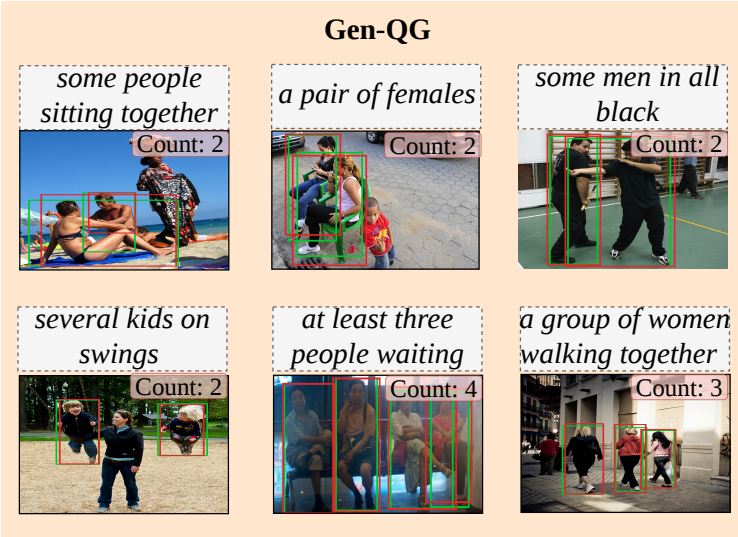


Figure 11: Qualitative results of Q-VG on Gen-QG, demonstrating quantitative grounding under generalized expressions. The red and green boxes are the predicted and ground truth ones, respectively. Count denotes the predicted number of referred objects.



Figure 12: Qualitative results of Q-VG on Comp-QG, demonstrating quantitative compositional grounding. The red and green boxes are the predicted and ground truth ones, respectively. Count denotes the predicted number of referred objects.

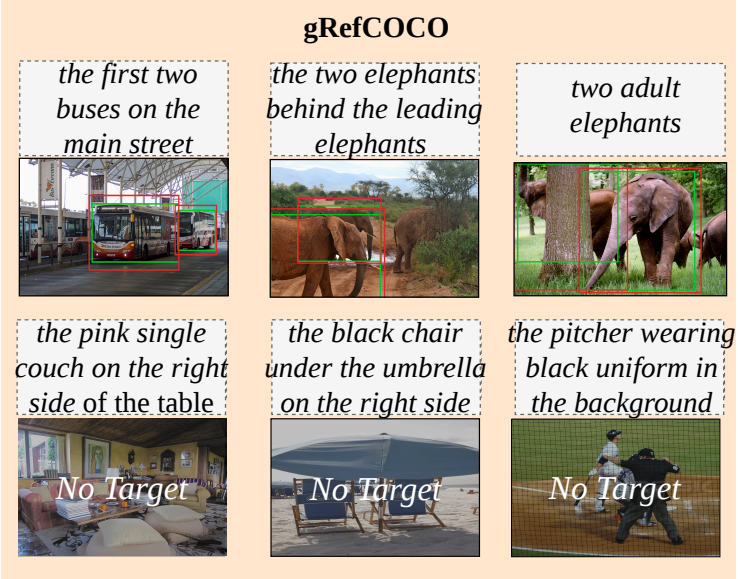


Figure 13: Qualitative results of Q-VG on gRefCOCO, demonstrating weakly supervised grounding. The red and green boxes are the predicted and ground truth ones, respectively.