Multi-label Instance-level Generalised Visual Grounding in Agriculture

Mohammadreza Haghighat[1,2⋆], Alzayat Saleh[1,2], Mostafa Rahimi Azghadi[1,2,3⋆]

[1] College of Science and Engineering, James Cook University, Townsville, QLD, Australia
[2] Centre for AI and Data Science Innovation, James Cook University, Townsville, QLD, Australia
[3] Australian Research Council Training Centre in Plant Biosecurity, Townsville, Australia
reza.haghighat@my.jcu.edu.au, alzayat.saleh@my.jcu.edu.au, mostafa.rahimiazghadi@jcu.edu.au

Fig. 1: Overview of the gRef-CW Dataset and Weed-VG Method. This figure summarises the paper's main contributions for tackling the challenges of visual grounding (VG) in agricultural imagery. It presents gRef-CW, a generalised Referring-expression dataset for Crop and Weed visual grounding, comprising 8,034 high-resolution field images and 82,592 annotations at both image-level and instance-level. The figure also highlights Weed-VG, a modular method that enables existence-aware instance grounding, first determining whether the referred instance is present and then localising it.

Abstract. Understanding field imagery such as detecting plants and distinguishing individual crop and weed instances, is a central challenge in precision agriculture. Despite progress in vision–language tasks like captioning and visual question answering, Visual Grounding (VG), localising language-referred objects, remains unexplored in agriculture. A key

⋆ Corresponding authors.

reason is the lack of suitable benchmark datasets for evaluating grounding models in field conditions, where many plants look highly similar, appear at multiple scales, and the referred target may be absent from the image. To address these limitations, we introduce gRef-CW, the first dataset designed for generalised visual grounding in agriculture, including negative expressions. Benchmarking current state-of-the-art grounding models on gRef-CW reveals a substantial domain gap, highlighting their inability to ground instances of crops and weeds. Motivated by these findings, we introduce Weed-VG, a modular framework that incorporates multi-label hierarchical relevance scoring and interpolation-driven regression. Weed-VG advances instance-level visual grounding and provides a clear baseline for developing VG methods in precision agriculture. Code and data are available at `https://github.com/MHaghighat98/WeedVG-gRefCW`.

Keywords: Precision Agriculture · Generalised Referring Expression Comprehension · Multi-modal Dataset · Hierarchical Multi-label Learning

1 Introduction

Agriculture is essential for sustaining human life, and effective management of crops, weeds, diseases, and pests is critical for stable production [1,2,38,46,51,66]. While VLMs have advanced tasks such as image captioning and Visual Question Answering (VQA) [9,31,53], VG—localising objects in images based on language queries—remains unexplored in agriculture [44, 56]. Unlike traditional object detectors, which are limited to predefined categories, the generalised form of VG (gVG) can handle multiple or absent targets and offer instance-aware localisation in complex field environments [16,27,28,45,50,57].

Crowded agricultural fields, with crops and weeds of varying sizes and absent targets, expose the limitations of standard VG models. Language-based queries offer flexible, context-aware specification, enabling monitoring of crop and weed presence, estimating quantities, and supporting selective weeding, fertilisation, irrigation, and harvesting. However, existing approaches [9,10,55] are still insufficient for real-world scenarios due to the domain shift. For instance, the well-known VG model GroundingDINO reports an mAP of 33.9 on agricultural images containing weeds [49]. Similarly, on the AgroBench dataset [51], open-source VLMs perform close to random for weed identification, with the best models achieving only 55.17% accuracy [51]. In addition, state-of-the-art grounding approaches such as InstanceVG [9] often fail to localise small objects. To systematically study these challenges, we introduce gRef-CW, the first agricultural VG dataset, designed with greater complexity and real-world variability compared to traditional VG benchmarks due to the following characteristics:

Novel and visually similar objects. In agricultural field images, VG is often challenged by objects that either fall outside the training distribution or closely resemble other instances [41]. Novel objects such as crops and weeds are unseen during training, and share overlapping features such as colour, shape, or texture [49]. Accurately distinguishing them is difficult when relying only on object-level features [15, 19, 34, 59, 62]. Therefore, effective detection must

incorporate instance-level information, such as position and size, that could be encoded in language descriptions.

Instances of varying sizes. Agricultural images contain a range of instances, varying in size from small seedlings to larger crops or weeds. Smaller objects occupy fewer pixels, making detection and accurate prediction of the bounding box more challenging [22, 23, 40]. Instance-level detection should be performed for all objects that are distinguishable by human experts.

To address the above challenges, gRef-CW is designed to support instance-level VG in agricultural field imagery by providing multi-level textual annotations. The dataset addresses the challenging task of distinguishing individual crops from weeds and determining their presence, considering significant intra-class variation. It contains 8,034 high-resolution images from CropAndWeed [52] with 78k crop and weed instances and 82k annotations, including 78,288 instance-level expressions and 4,304 image-level expressions (see Fig. 1). This comprehensive annotation enables the development and evaluation of VLMs in field images, making it suitable for gVG tasks. Aiming to improve gVG models' performance on gRef-CW, we also design an innovative framework to leverage a text-conditioned proposal generator and decompose gVG into hierarchical global existence detection and instance relevance through a multi-label constraint-enforcing approach. We conduct extensive experiments to thoroughly validate the effectiveness of our proposed model, demonstrating its superior performance.

2 Related Work

Vision-Language Models in Agriculture. In agriculture, VLMs have been applied to image captioning and VQA for applications including crop and plant health management as well as pest and insect detection [1, 17, 38, 46], However, their grounding capability remains unexplored in this domain, despite grounding being a core requirement for fine grained vision language understanding. Strengthening grounding can also improve downstream performance, including VQA, by enabling precise referring and localisation [47]. Existing agricultural multimodal benchmarks, including AgriBench [66], AgroBench [51], AgEval [1], CDDM [38], AgMMU [14], and MIRAGE [13], evaluate recognition and reasoning, but they are not designed for VG tasks. We address this gap by introducing a VG-specific agricultural dataset to leverage the power of VLMs in multimodal understanding, for instance-level crop and weed grounding.

Generalised Visual Grounding. gVG extends the classic VG task by addressing scenarios that involve zero or multiple targets. It can be further divided into two sub-tasks: generalised Referring Expression Comprehension (gREC)/Segmentation (gRES) [21, 32]. Models have been developed to improve the results mainly in scenarios where object classes are known but are distinguished using the referring expressions [3, 4, 6, 11, 63]. These methods leverage Transformer-based architectures to directly predict referred targets. Domain-specific VG faces challenges such as tiny entities and object similarity in aerial and remote sensing imagery [8,12,18,29,30,37,43,54,61,64,65] mainly due to the top-down perspective. Similarly, in medical imaging, grounding is difficult due to subtle boundaries and

reliance on expert semantic cues rather than appearance alone [7, 20, 24, 39, 67], making them distinct from natural scenes. Agricultural imagery presents analogous yet distinct challenges critical for gVG, where entity identification, such as detecting weeds and their growth stage, is complicated by high visual similarity to crops, specifically in crowded scenes. We leverage the gVG framework to address these fine-grained recognition challenges guided by language expressions, which remain unexplored in this domain.

3 Dataset

The gRef-CW dataset comprises field images representing diverse field scenes with varying crop and weed densities. It is derived from the CropOrWeed9 subset of the CropAndWeed dataset [52], with all labels mapped to eight crop types (Maize, Sugar beet, Bean, Pea, Sunflower, Soy, Potato, and Pumpkin) and one weed class. This subset was chosen because joint training across all crops with a unified weed category improves crop identification compared to single-crop models, enabling fine-grained crop–weed differentiation [52]. Bounding boxes and segmentation masks are sourced from CropAndWeed [52] where images were first pre-segmented into soil and vegetation via colour-based thresholding, later replaced by a CNN-based segmentation model. Bounding boxes were generated from these masks and manually refined, with ambiguous or densely populated images validated through multi-annotator voting to ensure consistency and quality. The dataset captures real field variability, where crops and weeds look alike, especially at early stages, and correctly determining weed presence or absence is crucial for effective field management.

Annotation pipeline. As shown in Fig. 2, we exclude instances smaller than 16×16 pixels, which are too small to be identified by humans. For each instance, we provide a referring expression that includes category, position, and size, along with a bounding box. Positions are extracted programmatically using a 3×3 grid based on bounding-box centre for each instance. Sizes are categorised into four levels—tiny, small, medium, and large—based on bounding-box area (Fig. 2).

Dataset statistics. To improve robustness in category-absent scenarios, we generate negative referring expressions at both the image level (see the right box at step 3 in Fig. 2) and instance level expressions in the test set (Step 4) for the gRef-CW. This follows two strategies, Replace and Swap [35, 58]. Replace substitutes category of an instance referring expression, whereas Swap interchanges two attributes within the same category (see step 4 in Fig. 2). To collect negative instance-level expressions in the test set, we randomly select one-third of the total 11,997 candidates for each change: category replacement, size swapping, and position swapping. This yields a total of 9,186 negative expressions (3,706 negative category, 3294 negative size, 2186 negative position), and the remaining are positive. All negatives are stored alongside positive expressions in a consistent annotation format, enabling evaluation under both object presence and absence conditions. The dataset is split into train, validation, and test sets with a ratio of 70:15:15, ensuring balanced representation of images containing only crops, only weeds, both, or neither. An overview of gRef-CW statistics is provided

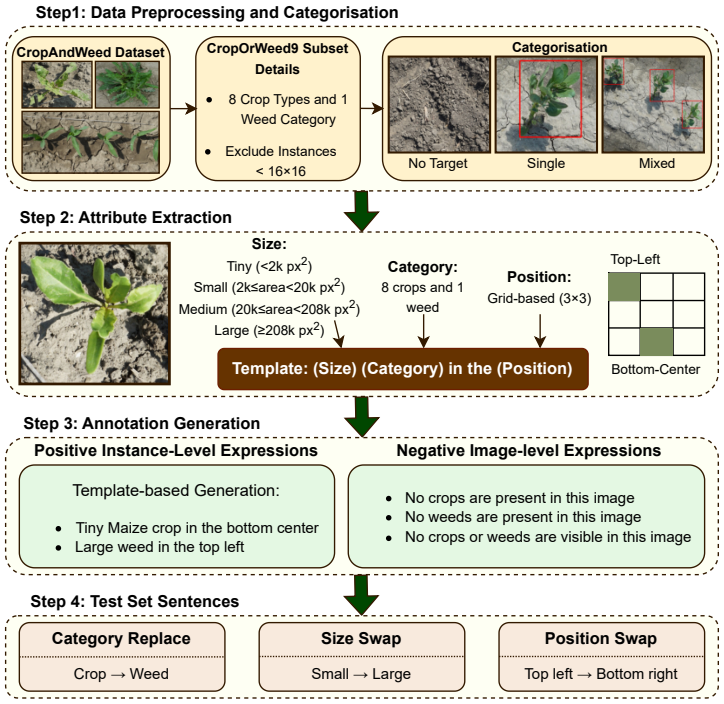

Fig. 2: gRef-CW Data Collection and Annotation Pipeline. The dataset is constructed through a four-stage process: (1) Filtering out instances unrecognisable by humans and categorising images as single, mixed, or no-target from the selected subset; (2) Extracting instance attributes (e.g. size, category, and position) (3) Composing attributes into natural language templates to generate positive instance-level referring expressions; negative image-level sentences include the non-existence and (4) Replacing categories or swapping attributes to create test sentences.

Dataset	General Statistics				Linguistics		Visual Properties		
	Images	Instances	Anns	Cats	Avg. Len.	Neg. Expr.	Inst. Size	Norm. Size	Img. Size
RefCOCO [60]	19,994	50,000	142,210	80	3.49	✗	230–640	0.17–1.0	640
RefCOCO+ [60]	19,992	49,856	141,564	80	3.58	✗	230–640	0.17–1.0	640
RefCOCOg [42]	26,711	54,822	104,560	80	8.46	✗	277–640	0.22–1.0	640
ReferItGame [26]	19,894	96,654	130,525	238	3.45	✗	360–480	$< 0.3^{\dagger}$	480
gRef-CW (Ours)	**8,034**	**78,288**	**82,592**	**9**	**6.34**	✓	**16–1,402**	**0.01–0.97**	**1,445**

† Minimum object size is not explicitly reported.

Table 1: Statistical comparison of classical referring expression datasets with gRef-CW. Instance and image dimensions are shown as square-root ranges. (Avg. Len. = average annotation length in words; Neg. Expr. = includes negative expressions.) This table contextualises the scale and complexity of gRef-CW rather than direct comparison of the annotation generation process, since our annotations are template-based, unlike the listed datasets.

in Table 1 compared to classical referring expression datasets. The Table shows gRef-CW's higher instance density, broader scale variation, instance-oriented expressions, and uniquely, the inclusion of negative expressions at both image and instance levels. As illustrated in Fig. 3, gRef-CW presents a challenging long-tailed distribution where 84.7% of instances are tiny or small, and scene complexity varies significantly from sparse to extremely dense clusters, where 30.7% of images contain more than 10 instances.

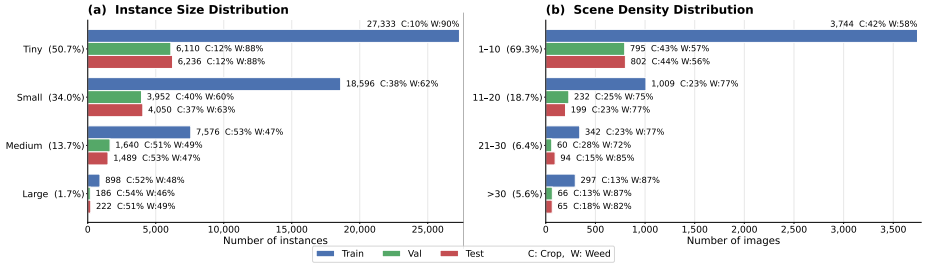

Fig. 3: Detailed instance distribution of the gRef-CW across splits. The figure displays (a) the count of instances stratified by scale (from Tiny to Large) and (b) the number of images categorised by scene density. Annotations within the bars indicate the percentage of Crop (C) and Weed (W) for each subset.

4 Method

The proposed Weed-VG architecture follows a modular design, where a grounding model generates proposals, which are then refined by the HRS module that captures both image- and instance-level information across two label levels. As shown in Fig. 4, given an image and a general text input to the grounding model, we first perform distance- and size-aware matching over the proposals for regression. Instance-level expressions are then aligned with their corresponding proposals (queries) and, together with the text representations, projected into a shared embedding space. The projected features are then processed by Multi-Head Cross-Attention (MHCA) followed by a Feed-Forward Network (FFN) to produce aligned multimodal relevance logits, through which instance relevance is learned contrastively in HRS during training. HRS decomposes relevance into two levels: Level-0 global existence detection, which predicts whether the text refers to any object in the image, and Level-1 instance relevance, which scores each candidate region. Word-level and sentence-level similarities are computed and combined via a learnable weight to form a unified referring score. A hierarchical constraint, enforced through a hierarchical multi-level loss, is applied to instance-level predictions by the global existence score, ensuring logical consistency between levels. This modular design allows the model to jointly predict whether a referent exists and which specific instances correspond to the query, while remaining compatible with standard grounding inference pipelines.

4.1 Hierarchical Relevance Scoring with Constraint Enforcement

Given a textual query and a set of candidate regions produced by the grounding model, HRS decomposes gVG into two levels. At Level 0 (*Existence Detection*), the

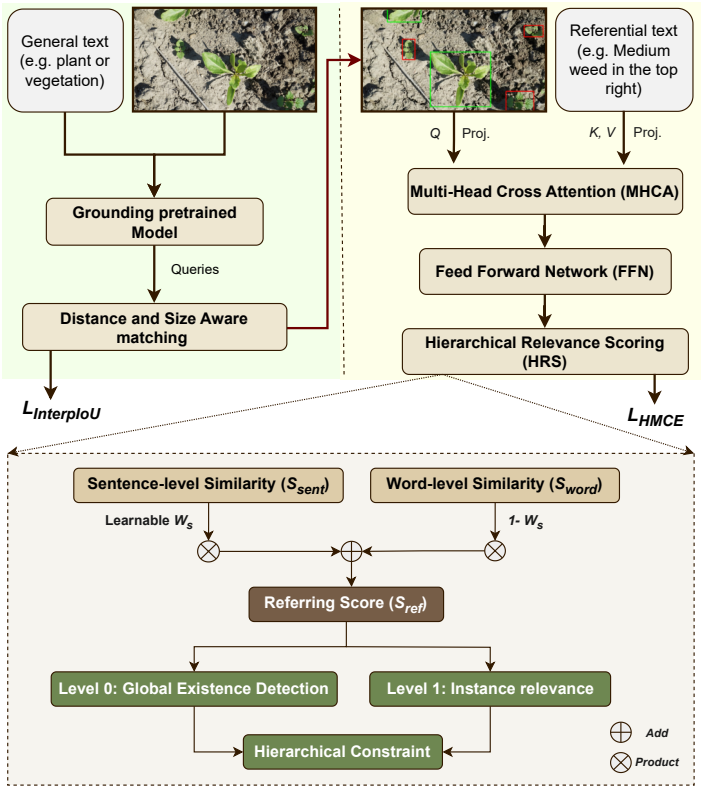

Fig. 4: Architecture of the Weed-VG framework. It extends a grounding model with an HRS module that fuses visual and textual features via Multi-Head Cross-Attention and an FFN. HRS decomposes relevance into two levels: (Level 0) global existence detection, which predicts whether the referred object appears in the image, and (Level 1) instance relevance, which ranks region proposals by integrating sentence-level and word-level similarities. A constraint enforces logical consistency by conditioning instance localisation on global existence.

model predicts whether the query refers to any instance in the image. At Level 1 (*Instance Discrimination*), it ranks candidate regions according to their relevance to the query. The final proposal score is obtained by enforcing a hierarchical constraint that aligns instance-level relevance with the global existence prediction.

Hierarchical Relevance Scoring. Our supervision strategy is structured across two hierarchical levels to ensure robust grounding. *Level 0: Existence Formulation.* At the coarsest granularity, the model determines the global semantic state of the image. This is formulated as a multiclass classification task over a fixed vocabulary of image-level sentences. We define the pooled level-0 logit k by selecting the maximum proposal score for a given sentence, and compute the resulting image-level class probability as

$$s_{\text{pool}}(k) = \max_j s(v_j, t_k), \quad P(k \mid I) = \text{Softmax}\big(s_{\text{pool}}(k)\big). \tag{1}$$

where v_i denotes the visual embedding of the i-th proposal and t_j denotes the textual embedding of the j-th sentence in the predefined level-0 vocabulary for the current image I. Optimisation at this stage employs a cross-entropy loss L_{lvl0}, ensuring the model verifies global existence before performing instance-level scoring. *Level 1: Instance Relevance Formulation.* Conditioned on the global existence prediction, the model proceeds to rank valid object instances. We optimise this ranking using a Binary Cross-Entropy (BCE) loss, L_{lvl1}, which accommodates both single-positive and multi-positive scenarios. The instance-level relevance logits are derived via a contrastive formulation that synergises sentence-level and word-level query-to-text similarities, which is made existence-aware by fusing with the level-0 logit k to produce S_{ref}.

Sentence-level textual features f_s are extracted by applying valid-mask max pooling over word-level embeddings f_t; this operation preserves the maximum response of each token across channels by selecting the highest value along the token dimension. Proposal-to-sentence similarity is then computed using a learnable temperature-scaled cosine similarity initialised to 0.07. Concurrently, the word-level branch computes a similarity matrix $S_{\text{word}} \in \mathbb{R}^{N \times N_t}$ between proposal features and individual word tokens.

To dynamically balance these contributions, a learnable weight w_s is generated by applying a multi-layer perceptron followed by a sigmoid activation to the global textual feature f_s. The referring score for each proposal is computed as:

$$S_{\text{ref}} = w_s \cdot S_{\text{sent}} + (1 - w_s) \cdot \text{MaxPool}(S_{\text{word}}). \tag{2}$$

This formulation adaptively integrates sentence-level semantics with fine-grained word-level cues, enabling robust instance discrimination.

Hierarchical Constraint Enforcement. A key logical dependency is that an instance cannot be localised if it does not exist in the image. We enforce this by applying a constraint on the instance-level loss using the image-level loss as a lower bound in the hierarchy: $L_{\text{lvl1}}^{\text{constrained}} = \max\left(L_{\text{lvl1}}, L_{\text{lvl0}}\right)$. This max operator guarantees that when global existence recognition is poor, instance discrimination cannot be performed. This effectively postpones L_{lvl1} until the existence prediction has stabilised, and leads to training dynamics that translate into a proposal-ranking behaviour at inference time.

4.2 IoU-Driven Interpolation for Bounding Box Regression

Crop and weed visual grounding poses challenges due to extreme scale variation depending on the growth stage, with object areas ranging from as little as 0.01% to 0.97% of the image. Direct application of standard IoU-based regression losses under such conditions often leads to unstable gradients and slow convergence [33]. To mitigate this issue, we adopt an IoU-driven interpolation strategy, termed InterpIoU, which stabilises bounding box optimisation without introducing handcrafted geometric penalties. Given a predicted bounding box B_{pred} and its corresponding ground truth B_{gt}, an intermediate box is constructed by linear interpolation,

$$B_{\text{int}} = (1 - \alpha)B_{\text{pred}} + \alpha B_{\text{gt}}, \qquad 0 < \alpha < 1. \tag{3}$$

The regression objective then combines the standard IoU loss with an auxiliary IoU term computed using the interpolated with $\alpha = 0.99$ to address extreme scale variation following the original implementation [33],

$$L_{\text{InterpIoU}}(B_{\text{pred}}, B_{\text{gt}}) = L_{\text{IoU}}(B_{\text{pred}}, B_{\text{gt}}) + L_{\text{IoU}}(B_{\text{int}}, B_{\text{gt}}), \tag{4}$$

where $L_{\text{IoU}} = 1 - \text{IoU}$. The auxiliary term provides smooth, non-zero gradients even when the predicted box is poorly aligned with the ground truth, effectively guiding predictions toward the target while preserving geometric integrity. This formulation is more robust to variations in plant instance scale and density across images.

Distance and Size Aware Matching. In the specific implementation, the matching cost is formulated to jointly account for overlap, centre distance, and relative scale differences between proposals and ground-truth boxes. This design encourages assignments that are consistent in position, size, and spatial alignment. The matching cost C_{ij} between the i-th proposal P_i and the j-th ground-truth instance G_j is defined as:

$$C_{ij} = (1 - \text{IoU}(P_i, G_j)) + \lambda_{\text{centre}} \|\mathbf{c}(P_i) - \mathbf{c}(G_j)\|^2 + \lambda_{\text{size}} \left(\frac{|w_{P_i} - w_{G_j}|}{w_{G_j}} + \frac{|h_{P_i} - h_{G_j}|}{h_{G_j}} \right) \tag{5}$$

where $\mathbf{c}(P_i)$ and $\mathbf{c}(G_j)$ denote the centres of the predicted and ground-truth boxes, respectively. In this equation, the first term measures the standard IoU overlap, the second term penalises the squared L_2 distance between box centres, and the third term captures the relative discrepancy in width and height. The weighting parameters are empirically chosen to prioritise box-centre alignment ($\lambda_{\text{centre}} = 2.0$) while remaining size-aware ($\lambda_{\text{size}} = 0.5$) due to extreme scale variation that causes the domination of centre error.

4.3 Loss Function

The final training objective integrates hierarchical semantic alignment and robust spatial localisation through a unified loss formulation:

$$L_{\text{total}} = L_{\text{HMCE}} + L_{\text{InterpIoU}}, \tag{6}$$

where L_{HMCE} is the Hierarchical Multi-label Constraint Enforcement loss, and $L_{\text{InterpIoU}}$ is the IoU-driven interpolation loss for bounding box regression.

Specifically, the hierarchical loss is defined as

$$L_{\text{HMCE}} = \lambda_0 \, L_{\text{lvl0}} + \lambda_1 \, L_{\text{lvl1}}^{\text{constrained}}. \tag{7}$$

The adaptive weights λ_0 and λ_1 are chosen based on image type: for multi-referent (mixed) images, $(1.0, 2.5)$ to prioritise instance discrimination; for single-category images, $(1.0, 2.0)$; and for empty images, $(1.0, 0)$ to focus solely on existence detection.

5 Experiments

5.1 Implementation Details

The proposed framework modularises grounding frameworks. Specifically, we employ GroundingDINO [36] with a Swin Transformer visual backbone and a

BERT-base text encoder in our experiments for proposal generation using the general text mentioned in Fig. 4. We then perform training in two stages to avoid instability observed in end-to-end joint training, which caused a reduction in regression performance. In the first stage, the model is initialised from the original checkpoint, and only the last decoder layer and box head are fine-tuned using the InterpIoU loss for 100 epochs. In the second stage, training is restricted to the projection/attention layers and the HRS module for 60 epochs using two-level labels, focusing on learning hierarchical semantics and instance-level distinctions. The epoch schedule is empirically selected to balance convergence and overfitting. Training batches are carefully constructed to include both positive instance-level expressions and negative sentences, including no-target and absence queries. A batch size of 4 is used, accounting for the four image types. Since annotations do not include negative sentences for images containing both crops and weeds, and hierarchical labels cannot be empty, we introduce a special token `[EMPTY]` to handle these cases. The maximum encoder input length for testing is set to 64 tokens. Optimisation is performed using AdamW with a cosine annealing learning rate schedule initialised at 2×10^{-4}. The data augmentation pipeline incorporates Copy–Paste augmentation to increase instance diversity, as well as random rotations and colour jittering to simulate variations in viewpoint, illumination, and plant appearance. All experiments were conducted using an NVIDIA A100 GPU on the Bunya HPC cluster and an NVIDIA RTX PRO 6000 Blackwell GPU. At inference, Weed-VG runs at 10 FPS (97.6 ms per image) and uses 1.5 GB of GPU memory.

5.2 Evaluation Metric

We evaluate four VG baselines on gRef-CW using Recall@0.5, Top-k Accuracy, and mIoU to assess retrieval, ranking, and localisation and to characterise dataset challenges. Since VLMs perform near random chance on agricultural tasks such as crop and weed detection [51], and grounding models cannot be fine-tuned on multi-level labels without HRS, direct comparison of the methodological improvements introduced by HRS with other methods is avoided to ensure fairness. Instead, the experiments aim to expose the domain gap via zero-shot benchmarking of existing models and establish Weed-VG as an improved agricultural baseline.

Due to the baselines' high instance miss on gRef-CW because of the sensitivity to threshold selection, we re-implement baselines to evaluate *all* their proposals given an instance sentence, and introduce Negative Accuracy (Neg-Acc), a threshold-free GIoU-based [48] metric, evaluated on sentences where no target exists. We report Neg-Acc alongside Recall to account for differences in proposal counts across baselines. A prediction is correct if its maximum GIoU with any ground truth is ≤ 0. Intuitively, Neg-Acc measures the model's ability to correctly refer to background. Evaluated together with Recall, it acts as a test of conditional understanding. Ideally, a model should both detect targets when present (High Recall) and abstain when absent (High Neg-Acc). Existing baselines fail this joint test: they achieve moderate detection only by flooding the scene with proposals, but their Neg-Acc collapse to $\sim$3–7.5% (SAM3 $\sim$25%).

Top-1 measures the fraction of queries where the highest-ranked proposal reaches IoU ≥ 0.5 with the ground-truth. High Top-5 but low Top-1 indicates the model finds plausible candidates but fails to rank the target correctly. To analyse language understanding in detail, metrics are further stratified by instance category, attribute type, and scene density (Tables 3 and 4).

5.3 Results

Main Benchmark Analysis. Table 2 illustrates baselines' dominant failure modes: weak instance ranking and localisation as well as insufficient recall. MDETR [25] achieves relatively high localisation quality, yet its Top-1 accuracy

Model	Validation				Test				
	Top-1	Top-5	R@0.5	mIoU	Top-1	Top-5	R@0.5	mIoU	Neg-Acc
MDETR [25]	10.20	12.81	7.82	52.35	10.16	12.97	7.78	54.19	3.32
GDINO-T [36]	12.89	33.04	21.50	18.93	11.92	31.99	20.34	17.44	2.88
GDINO-L [36]	19.74	42.58	28.16	23.08	20.38	43.49	28.73	23.68	7.52
SAM3 [5]	33.85	66.21	46.87	31.98	34.88	66.80	46.65	32.76	25.53
Weed-VG (Ours)	**63.01**	**81.64**	**55.71**	**58.12**	**62.42**	**82.45**	**55.44**	**57.25**	**78.35**

Table 2: Main results on the gRef-CW. We report Top-1, Top-5, R@0.5, and mIoU on both the Validation and Test sets. Negative Accuracy (Neg-Acc) is evaluated on the instance-level negative sentences provided in the Test set.

remains near 10% and R@0.5 below 8%. The large gap between mIoU (≈ 54) and R@0.5 (≈ 7.8) indicates that when a correct instance is retrieved, the box is spatially accurate; however, the model rarely retrieves the correct instance. In contrast, SAM3 [5] dramatically increases R@0.5 to 46.87% (Val) and 46.65% (Test), confirming higher coverage through dense proposal generation. However, its Top-1 accuracy remains limited to 33.85% / 34.88%. The 12-point gap between R@0.5 (≈ 46.7) and Top-1 (≈ 34.9) indicates that although the correct instance is often present among candidates, ranking fails in roughly one out of three cases. This behaviour is further reflected in mIoU (31.98 / 32.76), which remains substantially lower than MDETR [25] despite far higher recall, showing that proposal abundance alone does not guarantee precise localisation. Furthermore, increasing GroundingDINO backbone's capacity improves Test Top-1 accuracy, yet it still lags behind SAM3 [5], indicating that scaling the detector alone cannot resolve instance ambiguity. After integrating GroundingDINO [36] in Weed-VG, Top-1 accuracy increases to 63.01% (Val) and 62.42% (Test). Importantly, this improvement does not sacrifice recall: R@0.5 increases further to 55.71% / 55.44%. Simultaneously, mIoU rises to 58.12% / 57.25%. The joint increase across Top-1, R@0.5, and mIoU indicates that Weed-VG resolves both retrieval and ranking failures rather than improving a single metric dimension.

Scale-Specific Behaviour. Table 3 shows that the baselines exhibit severe collapse in metrics such as recall on tiny and small instances which are dominant in gRef-CW. On the Test set, both GDINO-T and GDINO-L [36] perform poorly on tiny instances, with Top-1 accuracies below 2%, highlighting failure under extreme scale reduction. SAM3 [5] improves tiny crop accuracy to 13.42%,

but this is still far below its 71.33% performance on large crops, revealing a substantial 58-point gap due to scale sensitivity. Weed-VG reduces this disparity substantially with the integration of GDINO-L. The scale gap between tiny and large crops shrinks to only 17 points (54.66 to 71.90). This confirms that hierarchical scoring stabilises ranking under extreme scale compression. Notably, on large crops GDINO-L achieves 74.33% Top-1, slightly higher than Weed-VG's 71.90%. However, the advantage of Weed-VG emerges in balancing the results across scales. This suggests that the contrastive relevance scoring, utilising both word- and sentence-level text processing while being existence-aware due to the hierarchy, highly benefits target referral.

Method	Set	Crop (Top-1 | mIoU)				Weed (Top-1 | mIoU)			
		Tiny	Small	Med	Large	Tiny	Small	Med	Large
MDETR [25]	Val	—†	4.17 | 22.84	10.28 | 52.31	11.29 | 60.29	—†	2.08 | 19.27	10.12 | 50.80	12.91 | 69.14
	Test	—†	3.31 | 21.18	10.34 | 53.56	11.71 | 63.40	—†	2.79 | 17.61	9.68 | 52.55	13.09 | 71.48
GDINO-T [36]	Val	0.68 | 0.74	11.83 | 14.87	23.36 | 35.38	35.98 | 60.04	0.00 | 0.26	6.50 | 9.03	18.29 | 28.72	43.18 | 73.82
	Test	0.00 | 0.05	7.56 | 9.82	23.75 | 34.86	31.64 | 53.53	0.31 | 0.72	6.45 | 8.75	20.79 | 31.48	43.32 | 68.73
GDINO-L [36]	Val	0.70 | 2.21	22.48 | 23.24	41.52 | 50.46	55.99 | **70.63**	0.75 | 1.29	12.44 | 13.85	37.15 | 45.26	63.53 | **79.55**
	Test	1.67 | 1.94	21.42 | 22.18	49.14 | 58.36	59.44 | **74.33**	0.83 | 1.53	12.27 | 13.70	37.30 | 44.94	65.20 | **82.30**
SAM3 [5]	Val	9.64 | 8.77	44.28 | 42.10	57.16 | 56.87	64.90 | 65.07	16.38 | 14.09	36.45 | 33.86	55.43 | 53.95	**76.85** | 77.52
	Test	13.42 | 11.84	48.63 | 46.58	61.20 | 60.38	71.33 | 70.58	16.17 | 13.97	35.57 | 32.53	55.12 | 53.43	**78.62** | 78.30
Weed-VG (Ours)	Val	**59.35** | **52.95**	**70.12** | **64.35**	**71.59** | **69.04**	**71.87** | 70.43	**57.25** | **50.44**	**59.18** | **55.28**	**67.93** | **65.70**	72.15 | 70.38
	Test	**54.66** | **50.28**	**64.78** | **61.33**	**70.92** | **66.16**	**71.90** | 68.70	**53.44** | **47.70**	**55.99** | **53.19**	**66.42** | **64.48**	76.79 | 70.69

† MDETR [25] produces no valid detections at this scale.

Table 3: Detailed performance breakdown by instance scale on the gRef-CW dataset. We report Top-1 accuracy and mIoU for crops and weeds separately on both Validation and Test sets. **Best results** are in bold, second-best are underlined.

Scene Density Robustness. Scene density further exposes ranking instability. As shown in Table 4, for crops in sparse scenes (1–10 instances), SAM3 [5] achieves 49.58 mIoU (Test). When density exceeds 30 instances, its mIoU drops to 25.81, a 23.77-point degradation. R@0.5 also decreases from 84.77 to 31.26, indicating severe interference among proposals. Robust performance in dense agricultural scenes, with more than 30 instances per image, is achieved by the proposed interpolation-based box regression enhanced with distance- and size-aware matching strategy in our framework, resulting in crop mIoU of 47.58 and weed mIoU remaining at 50.00.

Negative Expression Handling. As shown in Table 5, standard grounding pipelines tend to propose regions that have overlap with other ground-truth boxes as quantified by Neg-Acc in Table 5. SAM3 [5] improves Neg-Acc to 25.53%, largely due to proposal diversity, yet still fails in nearly three out of four negative cases. Dominant negative accuracy, with a consistent 78.35% average Neg-Acc across manipulation types, is achieved by the proposed Level-0 existence detection enforced by hierarchical constraints in our framework, resulting in the determination of referent existence and the generation of valid background proposals instead of referring to other ground-truth instances as shown in Fig. 5 (red box).

Scene Density	Crop (R@0.5 | mIoU)					Weed (R@0.5 | mIoU)				
	MDETR [25]	GDINO-T [36]	GDINO-L [36]	SAM3 [5]	Weed-VG (Ours)	MDETR [25]	GDINO-T [36]	GDINO-L [36]	SAM3 [5]	Weed-VG (Ours)
1–10 Instances										
Val	22.03 | 54.84	58.12 | 29.61	68.53 | 37.80	**85.48** | 44.72	84.26 | **66.80**	10.14 | 58.46	30.41 | 24.29	39.55 | 26.80	69.35 | 37.67	**69.73** | **58.75**
Test	21.94 | 57.76	51.09 | 24.37	67.93 | 38.67	**84.77** | 49.58	82.85 | **63.92**	10.83 | **59.77**	31.17 | 25.83	39.88 | 27.65	**69.37** | 37.63	69.02 | 56.36
11–20 Instances										
Val	14.85 | 48.33	40.86 | 18.37	51.43 | 29.00	67.70 | 42.07	**73.63** | **62.56**	3.75 | 46.43	12.00 | 10.56	18.68 | 12.33	41.23 | 23.85	**53.77** | **53.36**
Test	16.69 | 43.90	39.88 | 15.77	58.21 | 30.52	69.28 | 41.57	**74.59** | **60.35**	3.71 | **52.95**	11.48 | 11.93	18.19 | 13.32	39.84 | 22.37	**54.64** | 50.50
21–30 Instances										
Val	8.59 | 50.97	13.79 | 14.99	18.38 | 19.56	30.55 | 32.66	**44.15** | **58.66**	3.23 | 44.21	9.49 | 7.52	14.61 | 12.45	32.64 | 21.45	**47.91** | **54.58**
Test	12.50 | 54.92	28.89 | 13.54	38.69 | 29.54	52.08 | 41.51	**60.42** | **57.87**	3.52 | 49.55	9.07 | 8.64	15.23 | 11.15	31.13 | 18.92	**46.82** | **51.19**
>30 Instances										
Val	4.26 | 42.02	17.71 | 12.22	23.99 | 18.13	36.10 | 21.33	**50.67** | **53.26**	2.20 | 47.09	4.33 | 6.36	6.88 | 9.57	18.08 | 16.65	**32.10** | **50.11**
Test	4.75 | 40.33	12.80 | 9.09	20.11 | 16.40	31.26 | 25.81	**44.42** | **47.58**	0.94 | 41.58	4.08 | 5.15	7.53 | 8.60	18.93 | 19.57	**31.71** | **50.00**

Table 4: Scalability with respect to scene density on the gRef-CW dataset. We report R@0.5 and mIoU for crops and weeds on both Validation and Test sets, grouped by number of instances in the scene. **Best results** are in bold, second-best are underlined.

Qualitative Results. As shown in Fig. 5, the framework demonstrates proper instance recall and localisation. It also effectively rejects absent crop instances of the same size and spatial location by referring to background regions. Notably, when the category changes but the size and location remain similar, Weed-VG avoids making false positives by referring to the same instances (see bottom right, first image). However, limitations remain in highly ambiguous scenarios. The model occasionally suffers from partial recall as shown in the second image, top right. Furthermore, extreme visual similarity between early-stage "tiny" plants can cause category confusion, overriding both instance referral and the rejection mechanism.

Method	Neg-Acc			Avg.
	Replace	Swap		
	Category	Size	Position	
MDETR [25]	3.26	4.91	54.55	3.32
GDINO-T [36]	14.62	11.32	16.82	2.88
GDINO-L [36]	12.84	13.36	42.42	7.52
SAM3 [5]	41.02	34.03	60.75	25.53
Weed-VG (Ours)	**79.15**	**77.77**	**78.01**	**78.35**

Table 5: Negative Accuracy on the gRef-CW Test set. We report Neg-Acc for Replace and Swap changes, together with the weighted average across categories. **Best results** are in bold, second-best are underlined. Avg. denotes the average results reported, which are conditional on correct positive grounding.

Fig. 5: Qualitative results of Weed-VG on positive and negative referring expressions in gRef-CW across challenging scales (tiny–medium).

5.4 Ablations

To analyse the contribution of the components in the proposed framework, we conduct an ablation study as shown in Table 6. Using only sentence-level similarity drops Top-1 by 13.13 and mIoU by 7.9, while word-only cues degrade further to 36.19% Top-1, showing that grounding requires joint sentence semantics and token alignment. Removing Query Projection causes the largest collapse for Top-1 and mIoU, indicating that effective instance grounding highly depends on projecting visual and textual features into a shared learnable space that jointly captures the context of both image-level and instance-level expressions hierarchically. Removing the Hierarchical Constraint sharply reduces Neg-Acc (78.35 to 41.60) while localisation and Recall remain similar, demonstrating explicit encoding absence in the referring score without harming ranking. Finally, replacing InterpIoU coupled with the matching strategy, with the original GIoU implementation in GroundingDINO [36] drops mIoU by 9.53 and Top-1 consequently, particularly for smaller instances.

Method	Top-1	Top-5	R@0.5	mIoU	Neg-Acc
Sentence-only	49.29	74.06	50.53	49.35	74.12
Word-only	36.19	67.61	44.20	39.22	71.55
w/o Query Projection	33.20	58.02	38.43	35.39	69.83
w/o Hierarchical Constraint	59.87	80.57	53.87	55.83	41.60
w/o InterpIoU	49.15	73.28	44.61	47.72	75.88
Full Model	**62.42**	**82.45**	**55.44**	**57.25**	**78.35**

Table 6: Ablation study of the proposed framework. We evaluate the impact of different text processing strategies (Sentence-only vs. Word-only), removal of the Query Projection, removal of the Hierarchical Constraint, and the impact of the InterpIoU coupled with the matching strategy. All metrics are reported on the Test set.

6 Conclusion, Limitations and Future Work

We introduce gRef-CW, a real-world Crop and Weed dataset with 82k referring expressions, including negative sentences, providing clear value for benchmarking/training grounding frameworks across extreme scale variations, dense scenes, and different rejection scenarios. Evaluations of existing VG methods reveal substantial gaps in handling these challenges. To address this, we propose Weed-VG, a modular framework that is designed to integrate existing grounding models. Within Weed-VG, we introduce HRS for target existence and instance ranking, along with interpolation-driven regression to manage scale variation and high scene density. Experiments on standard grounding and negative-accuracy metrics demonstrate that explicitly modelling existence alongside instance ranking leads to more robust and reliable performance in crop and weed scenarios.

Limitations and Future Work. This work faces several constraints. Findings are limited by top-down image viewpoints, a reliance on initial proposal quality, and performance degradation in extremely dense scenes exceeding 30 instances. Furthermore, the dataset lacks broad agricultural category coverage and environmental diversity, while its template-based annotations restrict the

evaluation on free-form text, and richer contextual expressions. Applying WeedVG's modular, multi-level labelling scheme, which captures target non-existence and instance-level referrals, to standard benchmarks is possible but requires significant annotation effort. Future efforts will focus on exploring efficient annotation strategies, validating the HRS approach on broader visual grounding benchmarks, real-world deployment and developing larger-scale, agriculture-specific datasets.

Acknowledgement

Mostafa Rahimi Azghadi acknowledges support from the Australian Government through the Australian Research Council Training Centre in Plant Biosecurity (IC230100027).

References

1. Arshad, M.A., Jubery, T.Z., Roy, T., Nassiri, R., Singh, A.K., Singh, A., Hegde, C., Ganapathysubramanian, B., Balu, A., Krishnamurthy, A., et al.: Leveraging vision language models for specialized agricultural tasks. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6320–6329. IEEE (2025)
2. Azghadi, M.R., Olsen, A., Wood, J., Saleh, A., Calvert, B., Granshaw, T., Fillols, E., Philippa, B.: Precision robotic spot-spraying: Reducing herbicide use and enhancing environmental outcomes in sugarcane. Computers and Electronics in Agriculture **235**, 110365 (2025)
3. Bai, S., Li, M., Liu, Y., Tang, J., Zhang, H., Sun, L., Chu, X., Tang, Y.: Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. arXiv preprint arXiv:2505.14231 (2025)
4. Cai, Z., Ke, F., Jahangard, S., de la Banda, M.G., Haffari, R., Stuckey, P.J., Rezatofighi, H.: Naver: A neuro-symbolic compositional automaton for visual grounding with explicit logic reasoning. arXiv preprint arXiv:2502.00372 (2025)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Chen, Z., Wang, P., Ma, L., Wong, K.Y.K., Wu, Q.: Cops-ref: A new dataset and task on compositional referring expression comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10086–10095 (2020)
7. Chen, Z., Zhou, Y., Tran, A., Zhao, J., Wan, L., Ooi, G.S.K., Cheng, L.T.E., Thng, C.H., Xu, X., Liu, Y., et al.: Medical phrase grounding with region-phrase context contrastive alignment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 371–381. Springer (2023)
8. Choudhury, S., Kurkure, P., Banerjee, B.: Improving visual grounding in remote sensing images with adaptive modality guidance. ISPRS Journal of Photogrammetry and Remote Sensing **224**, 42–58 (2025)
9. Dai, M., Cheng, W., Liu, J.J., Yang, L., Feng, Z., Yang, W., Wang, J.: Improving generalized visual grounding with instance-aware joint learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

10. Dai, M., Cheng, W., Zhuang, J., Liu, J.j., Zhao, H., Feng, Z., Yang, W.: Propvg: End-to-end proposal-driven visual grounding with multi-granularity discrimination. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7058–7068 (2025)
11. Ding, H., Tang, S., He, S., Liu, C., Wu, Z., Jiang, Y.G.: Multimodal referring segmentation: A survey. arXiv preprint arXiv:2508.00265 (2025)
12. Ding, Y., Xu, H., Wang, D., Li, K., Tian, Y.: Visual selection and multistage reasoning for rsvg. IEEE Geoscience and Remote Sensing Letters **21**, 1–5 (2024)
13. Dongre, V., Gui, C., Garg, S., Nayyeri, H., Tur, G., Hakkani-Tür, D., Adve, V.S.: Mirage: A benchmark for multimodal information-seeking and reasoning in agricultural expert-guided conversations. arXiv preprint arXiv:2506.20100 (2025)
14. Gauba, A., Pi, I., Man, Y., Pang, Z., Adve, V.S., Wang, Y.X.: Agmmu: A comprehensive agricultural multimodal understanding and reasoning benchmark. arXiv preprint arXiv:2504.10568 (2025)
15. Guo, H., Zhu, J., Fan, W., Yi, C., Jiang, F.: Beyond object categories: Multi-attribute reference understanding for visual grounding. arXiv preprint arXiv:2503.19240 (2025)
16. Guo, K., Huang, Y., Jia, T., Song, X., Sun, S., Wei, H., Han, X.F., Huang, S., Strisciuglio, N., Li, S.: Visual grounding in 2d and 3d: A unified perspective and survey. Information Fusion p. 103625 (2025)
17. Haghighat, M., Saleh, A., Azghadi, M.R.: Multimodal language models in agriculture: A tutorial and survey. Information Fusion p. 104042 (2025)
18. Hang, R., Xu, S., Liu, Q.: A regionally indicated visual grounding network for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing (2024)
19. Hao, X., Zhu, K., Guo, H., Guo, H., Jiang, N., Lu, Q., Tang, M., Wang, J.: Referring expression instance retrieval and a strong end-to-end baseline. arXiv preprint arXiv:2506.18246 (2025)
20. He, J., Li, P., Liu, G., Zhong, S.: Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
21. He, S., Ding, H., Liu, C., Jiang, X.: Grec: Generalized referring expression comprehension. arXiv preprint arXiv:2308.16182 (2023)
22. Hemanthage, B., Bilen, H., Bartie, P., Dondrup, C., Lemon, O.: Recantformer: Referring expression comprehension with varying numbers of targets. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 21784–21798 (2024)
23. Hu, Y., Wang, Q., Shao, W., Xie, E., Li, Z., Han, J., Luo, P.: Beyond one-to-one: Rethinking the referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4067–4077 (2023)
24. Huy, T.D., Huynh, D.A., Xie, Y., Qi, Y., Chen, Q., Nguyen, P.L., Tran, S.K., Phung, S.L., Hengel, A.v.d., Liao, Z., et al.: Seeing the trees for the forest: Rethinking weakly-supervised medical visual grounding. arXiv preprint arXiv:2505.15123 (2025)
25. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: 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)
26. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)

27. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
28. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27831–27840 (2024)
29. Lan, M., Rong, F., Jiao, H., Gao, Z., Zhang, L.: Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–13 (2024)
30. Li, K., Wang, D., Xu, H., Zhong, H., Wang, C.: Language-guided progressive attention for visual grounding in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–13 (2024)
31. Li, X., Wen, C., Hu, Y., Yuan, Z., Zhu, X.X.: Vision-language models in remote sensing: Current progress and future trends. IEEE Geoscience and Remote Sensing Magazine **12**(2), 32–66 (2024)
32. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23592–23601 (2023)
33. Liu, H., Watanabe, H.: Interpiou: Robust bounding box regression loss within an interpolation-based iou framework. Neurocomputing p. 132230 (2025)
34. Liu, J., Chen, Q., Wang, Z., Tang, Y., Zhang, Y., Yan, C., Wang, D., Zhao, B., Li, X.: Aerialvg: A challenging benchmark for aerial visual grounding by exploring positional relations. arXiv preprint arXiv:2504.07836 (2025)
35. Liu, J., Yang, X., Li, W., Wang, P.: Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. arXiv preprint arXiv:2409.14750 (2024)
36. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
37. Liu, S., Ma, Y., Zhang, X., Wang, H., Ji, J., Sun, X., Ji, R.: Rotated multi-scale interaction network for referring remote sensing image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26658–26668 (2024)
38. Liu, X., Liu, Z., Hu, H., Chen, Z., Wang, K., Wang, K., Lian, S.: A multimodal benchmark dataset and model for crop disease diagnosis. In: European Conference on Computer Vision. pp. 157–170. Springer (2024)
39. Luo, L., Tang, B., Chen, X., Han, R., Chen, T.: Vividmed: Vision language model with versatile visual grounding for medicine. arXiv preprint arXiv:2410.12694 (2024)
40. Ma, T., Bai, B., Lin, H., Wang, H., Wang, Y., Luo, L., Fang, L.: When visual grounding meets gigapixel-level large-scale scenes: benchmark and approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22119–22128 (2024)
41. Madan, A., Peri, N., Kong, S., Ramanan, D.: Revisiting few-shot object detection with vision-language models. Advances in Neural Information Processing Systems **37**, 19547–19560 (2024)
42. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: European Conference on Computer Vision. pp. 792–807. Springer (2016)

43. Ou, R., Hu, Y., Zhang, F., Chen, J., Liu, Y.: Geopix: A multimodal large language model for pixel-level image understanding in remote sensing. IEEE Geoscience and Remote Sensing Magazine (2025)
44. Pantazopoulos, G., Özyiğit, E.B.: Towards understanding visual grounding in visual language models. arXiv preprint arXiv:2509.10345 (2025)
45. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: 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)
46. Quoc, K.N., Thu, L.L.T., Quach, L.D.: A vision-language foundation model for leaf disease identification. arXiv preprint arXiv:2505.07019 (2025)
47. Reich, D., Schultz, T.: Uncovering the full potential of visual grounding methods in vqa. arXiv preprint arXiv:2401.07803 (2024)
48. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
49. Robicheaux, P., Popov, M., Madan, A., Robinson, I., Nelson, J., Ramanan, D., Peri, N.: Roboflow100-vl: A multi-domain object detection benchmark for vision-language models. arXiv preprint arXiv:2505.20612 (2025)
50. Schulter, S., Suh, Y., Dafnis, K.M., Zhang, Z., Zhao, S., Metaxas, D., et al.: Omnilabel: A challenging benchmark for language-based object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11953–11962 (2023)
51. Shinoda, R., Inoue, N., Kataoka, H., Onishi, M., Ushiku, Y.: Agrobench: Vision-language model benchmark in agriculture. arXiv preprint arXiv:2507.20519 (2025)
52. Steininger, D., Trondl, A., Croonen, G., Simon, J., Widhalm, V.: The cropandweed dataset: A multi-modal learning approach for efficient crop and weed manipulation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3729–3738 (2023)
53. Tumu, A., Kordjamshidi, P.: Exploring spatial language grounding through referring expressions. arXiv preprint arXiv:2502.04359 (2025)
54. Wang, F., Wu, C., Wu, J., Wang, L., Li, C.: Multistage synergistic aggregation network for remote sensing visual grounding. IEEE Geoscience and Remote Sensing Letters **21**, 1–5 (2024)
55. Wang, Y., Ding, H., He, S., Jiang, X., Wei, B., Liu, J.: Hierarchical alignment-enhanced adaptive grounding network for generalized referring expression comprehension. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8042–8050 (2025)
56. Xiao, L., Yang, X., Lan, X., Wang, Y., Xu, C.: Towards visual grounding: A survey. arXiv preprint arXiv:2412.20206 (2024)
57. Xie, C., Zhang, Z., Wu, Y., Zhu, F., Zhao, R., Liang, S.: Described object detection: Liberating object detection with flexible expressions. Advances in Neural Information Processing Systems **36**, 79095–79107 (2023)
58. Yang, X., Liu, J., Wang, P., Wang, G., Yang, Y., Shen, H.T.: New dataset and methods for fine-grained compositional referring expression comprehension via specialist-mllm collaboration. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
59. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023)

60. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European conference on computer vision. pp. 69–85. Springer (2016)
61. Yuan, Z., Mou, L., Hua, Y., Zhu, X.X.: Rrsis: Referring remote sensing image segmentation. arXiv preprint arXiv:2306.08625 (2023)
62. Yuan, Z., Yan, X., Liao, Y., Zhang, R., Wang, S., Li, Z., Cui, S.: Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1791–1800 (2021)
63. Zeng, Y., Huang, Y., Zhang, J., Jie, Z., Chai, Z., Wang, L.: 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)
64. Zhan, Y., Xiong, Z., Yuan, Y.: Rsvg: Exploring data and models for visual grounding on remote sensing data. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–13 (2023)
65. Zhang, P., Zhang, Y., Wu, H., Liu, X., Hou, Y., Wang, L.: Language-guided object localization via refined spotting enhancement in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing (2025)
66. Zhou, Y., Ryo, M.: Agribench: A hierarchical agriculture benchmark for multimodal large language models. In: European Conference on Computer Vision. pp. 207–223. Springer (2024)
67. Zou, K., Bai, Y., Chen, Z., Zhou, Y., Chen, Y., Ren, K., Wang, M., Yuan, X., Shen, X., Fu, H.: Medrg: Medical report grounding with multi-modal large language model. arXiv preprint arXiv:2404.06798 (2024)