

RefDrone: A Challenging Benchmark for Referring Expression Comprehension in Drone Scenes

Zhichao Sun^{1,2}, Yepeng Liu¹, Zhiling Su¹, Huachao Zhu¹, Yuliang Gu¹, Yuda Zou¹, Zelong Liu¹
Gui-Song Xia³, Bo Du¹, Yongchao Xu^{1,2*}

¹School of Computer Science, Wuhan University ²Hubei LuoJia Laboratory

³School of Artificial Intelligence, Wuhan University

Abstract

Drones have become prevalent robotic platforms with significant potential in Embodied AI. A crucial capability for drone-based Embodied AI is Referring Expression Comprehension (REC), which enables locating objects with language expressions. Despite advances in REC for ground-level scenes, drones' unique capability for broad observation introduces distinct challenges: multiple potential targets, small-scale objects, and complex environmental contexts. To address these challenges, we introduce RefDrone, an REC benchmark for drone scenes. RefDrone reveals three key challenges: 1) multi-target and no-target scenarios; 2) multi-scale and small-scale target detection; 3) complex environments with rich contextual reasoning. To efficiently construct this dataset, we develop RDAnnotator, a semi-automated annotation framework where specialized modules and human annotators collaborate through feedback loops. RDAnnotator ensures high-quality contextual expressions while reducing annotation costs. Furthermore, we propose Number GroundingDINO (NGDINO), a novel method to handle multi-target and no-target cases. NGDINO explicitly estimates the number of objects referred to in the expression and incorporates this numerical pattern into the detection process. Comprehensive experiments with state-of-the-art REC methods demonstrate that NGDINO achieves superior performance on RefDrone, as well as on the general-domain gRefCOCO and remote sensing RSVG benchmarks.

1. Introduction

Drones/UAVs have become increasingly prevalent in personal and professional applications, including entertainment, package delivery, traffic surveillance, and emergency rescue [1, 45, 51]. Their mobility and broad observational capabilities make them promising platforms for Embodied

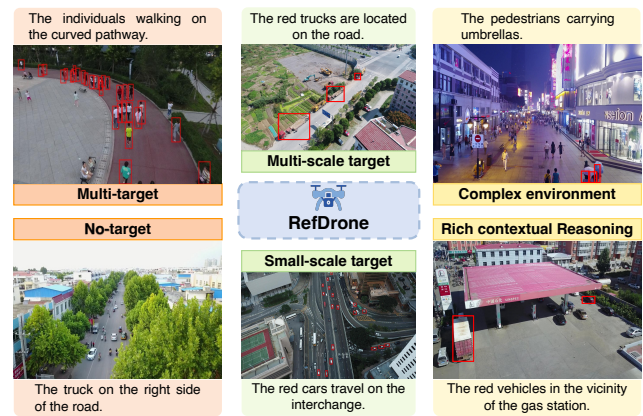


Figure 1. Examples of the various challenges in RefDrone dataset.

AI [13, 15, 25, 34, 37]. Notably, AerialVLN [37] is a pioneering work in drone-based Embodied AI, demonstrating the potential of integrating language instructions with aerial vision for navigation. To support such Embodied AI systems, a fundamental capability is Referring Expression Comprehension (REC) [5, 46, 53, 54], which bridges language understanding and visual perception by localizing specific objects in images based on textual descriptions. However, existing REC datasets predominantly adopt ground-level perspectives (e.g., RefCOCO [61]), leaving drone-based scenarios underexplored despite their unique challenges.

In this work, we introduce **RefDrone**, a challenging REC benchmark for drone scenes. RefDrone comprises 17,900 referring expressions annotated across 8,536 images, with 63,679 object instances. As illustrated in Figure 1, RefDrone presents three primary challenges: (1) **multi-target and no-target samples**, where expressions may refer to any number of objects (0 to 242); (2) **multi-scale and small-scale target detection**, featuring 31% small objects and 14% large objects; and (3) **complex environment with rich contextual reasoning**, encompassing diverse viewpoints, lighting conditions, intricate backgrounds, and rich descriptions of spatial relations, object attributes, and inter-object interac-

*Corresponding author: yongchao.xu@whu.edu.cn

Code available: github.com/sunzc-sunny/refdrone.git

Table 1. Comparison of REC datasets relevant to RefDrone. Avg. objects: objects per expression. Avg. length: words per expression. No target: expressions without referred objects. Expression type: expression generation method. Small target: percentage of small-scale objects.

	RefCOCO+/g [39, 61]	gRefCOCO [30]	D ³ [58]	RIS-CQ [20]	RSVG [63]	RefDrone (Ours)
Image source	COCO [28]	COCO	COCO	VG [24]+COCO	DIOR [26]	VisDrone [67]
Avg. objects	1.0	1.4	1.3	3.6	2.2	3.8
Avg. length	3.6/3.5/8.4	4.9	6.3	13.2	7.5	9.0
No target	✗	✓	✓	✗	✗	✓
Expression type	Manual	Manual	Manual	LLM	Templated	LVLm
Small target	0/0/0%	0.1%	6.3%	-	17.2%	31.1%

tions. As shown in Table 1, RefDrone offers greater diversity and complexity than existing REC benchmarks. We evaluate **26 representative REC models** including: specialized REC methods, task-specific LVLms, general LVLms, and closed-source API models, using zero-shot settings. All models exhibit poorer performance on RefDrone compared to standard REC datasets (*e.g.*, Qwen2.5-VL-7B [4] yields 26.52% Acc_{img} on RefDrone vs. 92.5% Acc_{img} on RefCOCO_{testA}), highlighting the inherent difficulty and unique challenges.

To enable efficient dataset construction, we develop RDAannotator, a semi-automated annotation pipeline for referring expression annotation in drone scenes. RDAannotator leverages multiple specialized LVLm-based modules, which collaborate within a feedback loop to generate and validate annotations. By reducing human involvement to quality control and minor refinements, RDAannotator achieves a cost of merely **\$0.0539 per expression** (GPT-4o API) and reduces human annotation time to under **one minute per expression**. This cost-efficiency makes RDAannotator a scalable solution for large-scale dataset construction and can be readily adapted to other REC tasks.

Furthermore, we propose Number GroundingDINO (NGDINO) to address multi-target and no-target challenges. Our key insight is that ***explicitly modeling the number of referred objects significantly enhances handling of these scenarios***. NGDINO includes three components: (1) a number prediction head estimating target object counts, (2) learnable number-queries capturing numerical patterns for varying object quantities, and (3) a number cross-attention module fusing number queries with detection queries for enhanced localization. Extensive experiments on our RefDrone benchmark, as well as the general-domain gRefCOCO [30] and remote sensing RSVG [63] datasets, demonstrate that NGDINO achieves substantial improvements, particularly in multi-target and no-target cases.

In summary, our contributions are listed as follows:

- **RefDrone Benchmark:** We introduce RefDrone, a comprehensive REC benchmark for drone scenes featuring three key challenges: multi-target/no-target scenarios, multi-scale/small-object detection, and complex contextual reasoning. We provide thorough evaluations of 26 representative REC models.

- **RDAannotator Framework:** We propose RDAannotator, a cost-effective semi-automated annotation framework that significantly reduces human effort while ensuring high-quality annotations. This framework is scalable for large-scale dataset construction and generalizable beyond drone imagery.
- **NGDINO Method:** We develop NGDINO, a novel approach explicitly modeling object counts to address multi-target and no-target cases. NGDINO achieves state-of-the-art performance on RefDrone with consistent improvements on gRefCOCO and RSVG.

2. Related Work

2.1. Referring expression understanding datasets

Referring expression understanding identifies visual regions corresponding to natural language expressions. Subtasks include Referring Expression Comprehension (REC), which predicts bounding boxes, and Referring Expression Segmentation (RES), which generates pixel-level masks. Early benchmarks such as ReferIt [23] and RefCOCO [61] were pioneering but mostly limited to single-target scenarios. Subsequent works introduced greater complexity. For instance, gRefCOCO [30] expanded to multi-target expressions, while D³ [58] and RIS-CQ [20] provided more descriptive and challenging language. However, these datasets typically involve a small, fixed number of targets per expression.

Domain-specific extensions have also emerged, including video action recognition (RAVAR [40]) and affordance detection (RIO [42]). In the aerial domain, RSVG [63] focuses on remote sensing data but uses expressions with limited contextual richness. AerialVG [33] focuses on using spatial relations for a single target, leaving the challenge of multi-target and no-target tasks. To fill this gap, our RefDrone benchmark provides a more complex and realistic setting by introducing complex multi-target scenarios with contextually rich expressions. Table 1 provides a detailed comparison with related benchmarks.

2.2. Referring expression comprehension methods

REC methods can be broadly categorized into large vision language models (LVLms) and specialist models.

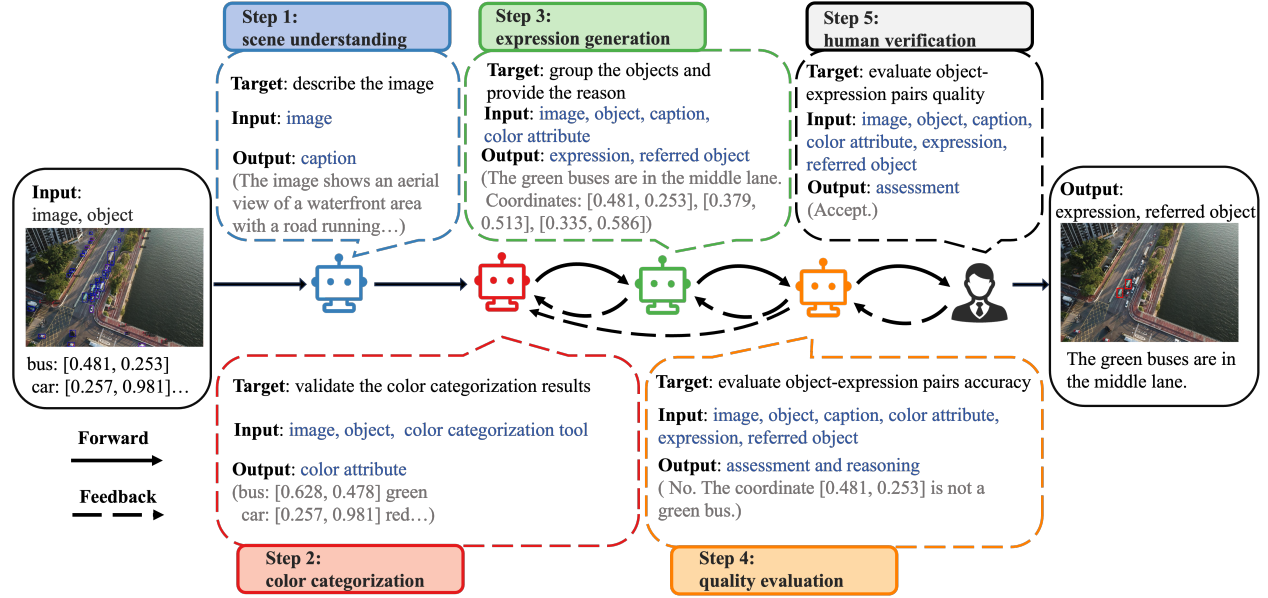


Figure 2. The overview of the RefDrone annotation process with RDAnnotator. Multiple specialized LVLM-based modules collaborate both with each other and human annotators through iterative feedback loops to generate high-quality annotations.

LVLMS [3, 6, 8, 21, 29, 41, 55, 60, 64] have recently been applied to REC tasks as part of evaluating their broader visual-language understanding capabilities. These models leverage extensive referring instruction tuning data to achieve competitive performance without task-specific architectural designs. To manage computational demands, LVLMS typically process downsampled images [32] or employ visual token reduction strategies [3, 57], which limits their ability to preserve fine-grained details. This constraint makes them less effective for grounding small objects, a crucial requirement in applications like aerial image analysis.

Specialist models include two-stage and one-stage methods. Two-stage methods [17–19, 31, 47] typically approach REC as a ranking task: first generating region proposals through an object detector, then ranking proposals based on language-vision alignment. Despite achieving strong accuracy, two-stage pipelines suffer from slow inference. In contrast, one-stage methods [14, 22, 27, 35, 36, 59] directly predict target regions guided by language input. These approaches leverage transformers to enable cross-modal interactions between visual and textual features. Among these, GroundingDINO (GDINO) [36] has gained widespread attention for its impressive results in REC tasks. Our work extends GDINO by incorporating explicit number modeling to handle multi-target and no-target scenarios, which are critical challenges in drone-based REC tasks.

3. RefDrone benchmark

3.1. Data source

The RefDrone benchmark is built upon VisDrone2019-DET [67], a high-quality drone-captured object detection

dataset. Images are collected across scenarios, illumination conditions, and flying altitudes. To ensure sufficient complexity for referring expression annotation, we filtered the dataset, retaining only images with at least three objects and excluding objects with bounding box areas smaller than 64 pixels. VisDrone2019-DET provides object categories and bounding box coordinates, which we convert to normalized center points (range 0-1). This approach reduces the token count for LVLMS while preserving spatial relationships.

3.2. RDAnnotator for semi-automated annotation

To construct our benchmark, we introduce RDAnnotator, a semi-automated annotation framework that balances annotation quality and scalability by integrating LVLM modules with a human-in-the-loop validation process. As illustrated in Figure 2, the framework operates in five steps.

Step 1: scene understanding. A scene-parsing module, powered by GPT-4o, generates three diverse captions for each image. These captions establish a foundational context by detailing spatial arrangements and object relationships for subsequent referring expression generation.

Step 2: color categorization. To identify color attributes, which are crucial discriminative features in REC, we employ a hybrid pipeline. This pipeline combines a CNN-based classifier (WideResNet-101 [62]) with high-level LVLM-based semantic reasoning to ensure accurate color attribution despite challenging lighting and atmospheric conditions.

Step 3: expression generation. We reformulate expression generation as an *object grouping task*. This module clusters semantically related objects and generates a textual justification for each grouping, which then serves as the referring expression. A dynamic feedback loop connects this

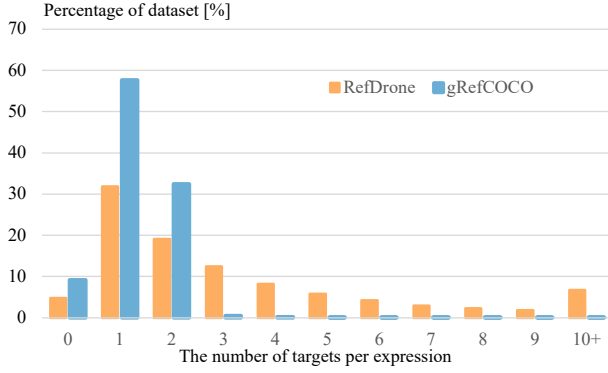


Figure 3. Object number distribution per expression in gRefCOCO [30] and RefDrone datasets.

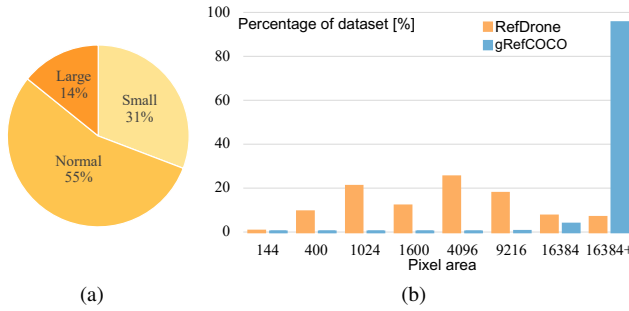


Figure 4. Object size distribution analysis. (a) Object size distribution in RefDrone dataset (small: $< 32^2 = 1024$ pixels, normal: 1024 to 9216 pixels, large: $> 96^2 = 9216$ pixels). (b) Object size histograms in RefDrone and gRefCOCO [30] datasets.

step to color categorization (Step 2), triggering re-analysis if novel color terms are generated, thereby ensuring attribute consistency across the dataset.

Step 4: quality evaluation. A validation module automatically assesses each generated object-expression pair for semantic accuracy and uniqueness. Correct annotations advance to human verification (Step 5). Incorrect annotations are routed back with targeted feedback: semantic errors are returned to expression generation (Step 3), while color inaccuracies are sent back to color categorization (Step 2).

Step 5: human verification. All generated annotations undergo a final human review, categorized into three tiers:

- *Direct acceptance.* Annotations satisfying all criteria are approved for the final dataset.
- *Refinement required.* Annotations with minor errors are corrected through human editing.
- *Significant issues.* Annotations with poorly grounded or inconsistent content trigger full regeneration.

This multi-level feedback system ensures high annotation fidelity. Expressions that consistently fail verification are designated as “no-target” samples, creating a set of contextually valid negative examples.

Each step leverages LVLMs via in-context learning with carefully designed task-specific prompts and examples (see

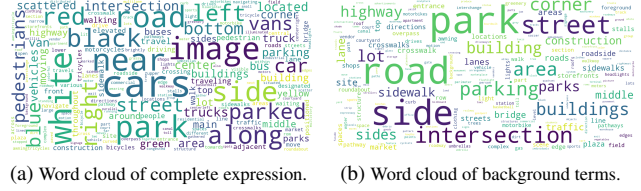


Figure 5. Word frequency visualization in RefDrone dataset.

Appendix). The effectiveness of our framework is demonstrated by the annotation outcomes: 42% of samples were directly accepted, 47% required minor refinement, and only 11% were rejected for re-annotation. Ultimately, RDAnnotator reduces human annotation effort by 85%, decreasing the average time per expression from 7 minutes to 1 minute. This efficiency is achieved at a low API cost of \$0.0539 per expression, demonstrating the framework’s scalability for creating large-scale, high-fidelity REC datasets.

3.3. Dataset analysis

The RefDrone dataset comprises 17,900 referring expressions for 63,679 object instances across 8,536 images, spanning 10 categories. The dataset preserves the original train, validation, and test splits from VisDrone2019-DET [67]. On average, each expression is 9.0 words long and refers to 3.8 objects. RefDrone is characterized by three primary challenges, illustrated in Figure 1:

1) Multi-target and no-target samples. In contrast to datasets like RefCOCO [61] that focus on single-object references, RefDrone features a significant portion of complex queries, including 11,362 multi-target and 847 no-target expressions. The number of targets per expression ranges from 0 to 242. As shown in Figure 3, this distribution presents a greater challenge than that of gRefCOCO [30], where expressions typically refer to only one or two objects.

2) Multi-scale and small-scale target detection. The dataset exhibits a wide distribution of object scales (Figure 4): small objects ($< 32^2$ pixels) account for 31%, medium objects (32^2 - 96^2 pixels) for 55%, and large objects ($> 96^2$ pixels) for 14%. The high variance in object scales, particularly the prevalence of small objects, underscores multi-scale and small-scale target detection challenges.

3) Complex environment with rich contextual reasoning. Images are captured in complex environments with diverse viewpoints, lighting, and dense backgrounds. Consequently, the referring expressions extend beyond simple attributes (e.g., color) and spatial relationships (e.g., ‘left of’) to describe complex object-object interactions (e.g., ‘the white trucks carrying livestock’) and object-environment interactions (e.g., ‘the white cars line up at the intersection’). This complexity is visualized in the word clouds in Figure 5.

3.4. Comparison to existing datasets

Table 1 situates RefDrone among existing REC datasets. Key distinguishing features of RefDrone are its high average

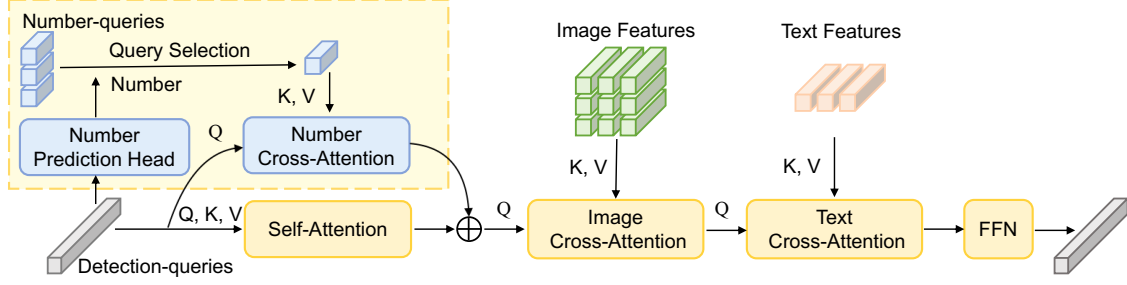


Figure 6. Architecture of a single decoder layer in Number GroundingDINO. Key modifications from GDINO [36] (highlighted in the yellow box) include: (1) a number prediction head (FFN) to estimate target count, (2) number-queries selected through the predicted number, (3) a number cross-attention between selected number-queries and detection-queries.

number of targets per expression and the use of an LVLM annotation pipeline. This pipeline generates expressions with richer contextual details compared to those from template-based methods or human-only annotation [66]. Although RIS-CQ [20] also uses an LLM for generation, its process is decoupled from visual content, leading to expressions that can be linguistically complex but visually ambiguous. While RSVG [63] targets small objects, the descriptive quality of its expressions is limited. In contrast, RefDrone comprehensively integrates these challenges, establishing it as a challenging benchmark in REC tasks.

3.5. Evaluation metrics

To properly assess performance on multi-target expressions, we introduce instance-level metrics alongside traditional image-level ones. Standard image-level metrics, which treat the entire set of predictions for an expression as a single entity, are insufficient for our task. They cannot granularly penalize a model for missing individual objects within a large group. Our proposed instance-level metrics address this limitation by providing a fine-grained assessment of a model’s ability to localize each individual target.

Instance-level metrics $\text{Acc}_{\text{inst.}}$ and $\text{F1}_{\text{inst.}}$: These metrics evaluate performance at the individual object level. We match each predicted bounding box to a ground-truth (GT) box. A prediction with an $\text{IoU} \geq 0.5$ with a GT box is a true positive (TP). Unmatched predictions are false positives (FP), and unmatched GT boxes are false negatives (FN). For “no-target” samples, the absence of any prediction is a true negative (TN), while any prediction is an FP. Accuracy and F1-score are then computed as:

$$\text{Acc}_{\text{inst.}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \quad (1)$$

$$\text{F1}_{\text{inst.}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}. \quad (2)$$

Image-level metrics $\text{Acc}_{\text{img.}}$ and $\text{F1}_{\text{img.}}$: These metrics assess performance on the entire expression, enforcing a stricter success criterion. For a given expression, a prediction is a true positive (TP) only if the set of predicted boxes

perfectly matches the set of ground-truth boxes. Any mismatch (e.g., missing, extra, or inaccurate boxes) results in a false positive (FP). TN and FP for “no-target” samples are defined as above. The metrics are calculated using the same formulas but on an image-wide basis.

4. NGDINO

We introduce Number GroundingDINO (NGDINO), a novel architecture designed to address the challenges of multi-target and no-target referring expression comprehension (REC). Our core insight is that explicit numerical reasoning about target counts is crucial for accurately grounding such expressions. NGDINO builds upon the strong foundation of GDINO [36], inheriting its dual-encoder, single-decoder structure. Our contributions are concentrated within the decoder, as highlighted in Figure 6, where we introduce three key components: (1) a number prediction head, (2) learnable number-queries with number-guided query selection, and (3) a number cross-attention module.

Number prediction head. To explicitly reason about the number of targets, we append a lightweight number prediction head to the decoder. This head takes the detection queries Q_{det} as input, processes them through a Feed-Forward Network (FFN), and applies mean pooling to produce a probability distribution over target counts:

$$N_{\text{prob}} = \text{softmax}(\text{MeanPool}(\text{FFN}(Q_{\text{det}}))), \quad (3)$$

$$N_{\text{pred}} = \text{argmax}(N_{\text{prob}}), \quad (4)$$

where $Q_{\text{det}} \in \mathbb{R}^{B \times L_d \times D}$ are the detection queries for a batch size B , query length L_d , and feature dimension D . To manage the long-tailed, Zipfian-like distribution of target counts [2], we discretize the output space into five bins: $\mathcal{C} = \{0, 1, 2, 3, 4+\}$, where $4+$ represents four or more targets. This approach transforms the open-ended regression problem into a stable multi-class classification task.

Number-guided query selection. To inject numerical priors into the decoding process, we introduce learnable number-queries, $Q_{\text{num}} \in \mathbb{R}^{B \times L_n \times D}$, where L_n is the length of

number-queries. These queries are trained to encode distinct numerical patterns. The predicted number, N_{pred} , then acts as an index to select a dedicated slice of these queries:

$$Q_{num}^{sel} = Q_{num}[:, L_s \cdot N_{pred} : L_s \cdot (N_{pred} + 1), :], \quad (5)$$

where L_s is the length of number-queries per category. This mechanism creates a mapping between the predicted cardinality and queries encoding patterns.

Number cross-attention module. The selected number-queries, Q_{num}^{sel} , are fused with the detection queries, Q_{det} , via a number cross-attention module. This module operates in parallel with the self-attention layer within the decoder. The detection queries serve as the query (Q), while the selected number-queries serve as both key (K) and value (V). The output of this module is added to the output of the self-attention layer, enriching the detection queries with numerical context before they are passed to subsequent layers.

Training objective. The model is trained end-to-end by augmenting the GDINO losses (for bounding box regression and label classification) with a Cross-Entropy loss for the number prediction task. Hyperparameters, including the selected number-query length ($L_s = 10$) and total number-query length ($L_n = 50$, representing 5 categories $\times$ 10 queries), were determined via ablation studies in the Appendix.

5. Experiments

We establish a comprehensive benchmark comprising 26 representative methods capable of performing REC tasks: 3 specialized REC methods, 7 task-specific LVLMs for REC, 12 general LVLMs with varying parameter scales, and 4 closed-source API models. To evaluate our proposed NGDINO, we conduct experiments on our RefDrone dataset and two public benchmarks: gRefCOCO [30] and RSVG [63].

5.1. Zero-shot results

As presented in Table 2, we evaluate all models in a zero-shot setting to assess their generalization to the drone scenes.

Overall performance. Among all models, the open-source Qwen3-VL (235B) [49] achieves state-of-the-art performance, with 58.79% $F1_{inst.}$, 41.93% $Acc_{inst.}$, 52.16% $F1_{img.}$, and 36.89% $Acc_{img.}$. These results surpass not only other open-source models but also the specialized closed-source API, DINO-XSeek [44] (54.47% $F1_{inst.}$), which is explicitly designed for REC tasks.

Analysis of specialized methods. Specialized REC models (e.g., MDETR [22], GLIP [27], GroundingDINO [36]) show limited zero-shot transfer to RefDrone, with $F1_{inst.}$ generally below 10%. This likely stems from pre-training on general-domain datasets, which restricts their cross-domain adaptability. Task-specific LVLMs for REC show varied performance. Models limited to predicting a single bounding box,

like Shikra [8], ONE-PEACE [52], and SPHINX-v2 [29], fail in multi-target scenarios, resulting in $F1_{inst.}$ scores below 2%. In contrast, recent Rex-Omni [21] achieves competitive performance (54.06% $F1_{inst.}$ / 43.90% $F1_{img.}$), approaching the state-of-the-art.

Analysis of general LVLMs. A strong correlation is observed between model scale and zero-shot REC performance in general-purpose LVLMs. For instance, the Qwen3-VL series shows consistent improvement with scale: the 4B model achieves 48.68% $F1_{inst.}$, which increases to 51.66% for the 8B model and 56.88% for the 30B model. This scaling trend suggests that larger model capacity directly enhances localization abilities without task-specific tuning. A notable exception is DeepSeek-VL2 [57], where the small variant achieves better performance.

Performance of closed-source APIs. Commercial models do not consistently outperform open-source alternatives. DINO-XSeek [44] achieves 54.47% $F1_{inst.}$, lagging behind Qwen3-VL-235B. Other prominent APIs, including Gemini 2.5 Pro, yield low scores ($F1_{inst.} < 4\%$). Furthermore, models from the GPT and Claude families failed to produce outputs in the required bounding-box format. This disparity indicates that many leading commercial models are not optimized for fine-grained localization tasks like REC.

5.2. Fine-tuning results

Performance on the RefDrone Dataset. Table 3 reports the fine-tuning performance of specialized REC methods for 50 epochs on the RefDrone dataset. Our proposed NGDINO shows consistent gains over the GDINO baseline. Specifically, the Swin-Tiny model NGDINO-T surpasses GDINO-T by 3.47%, 3.97%, 1.97%, and 1.57% in $F1_{inst.}$, $Acc_{inst.}$, $F1_{img.}$, and $Acc_{img.}$, respectively. NGDINO-B achieves the best performance, with 72.51% $F1_{inst.}$, 57.22% $Acc_{inst.}$, 57.84% $F1_{img.}$, and 42.54% $Acc_{img.}$, outperforming the state-of-the-art LVLM Qwen3-VL-235B by a significant margin. Moreover, NGDINO uses orders of magnitude fewer parameters than LVLMs, making it more suitable for deployment on resource-constrained edge platforms such as drones.

Performance on public benchmarks. To validate the effectiveness of our improvements over GDINO in multi-target scenarios, we evaluate NGDINO on two benchmarks containing multi-target samples: gRefCOCO [30] and RSVG [63]. Models are fine-tuned for 5 epochs on both benchmarks. Table 4 presents results on gRefCOCO [30], which includes both multi-target and no-target samples. We adopt two metrics: $Pr@0.5$, measuring the percentage of predictions achieving $F1 = 1$ at $IoU \geq 0.5$, and N-acc, denoting accuracy on no-target samples. NGDINO-T demonstrates improvements over GDINO-T, with $Pr@0.5$ gains of 0.8% and 0.3% on test A and test B, respectively. The improvements are more modest, which can be attributed to the relatively sim-

Table 2. Experimental results of zero-shot baselines on RefDrone benchmark. The best results in each group are denoted with **bold**.

Categories	Methods	Params	Release Date	F1 _{inst.}	Acc _{inst.}	F1 _{img.}	Acc _{img.}
Specialized REC Methods	MDETR _{ResNet101} [22]	0.19B	2021/04/26	8.42	4.41	2.99	1.63
	GLIP _{Swin-Tiny} [27]	0.15B	2021/12/06	5.46	3.84	9.20	8.54
	GroundingDINO _{Swin-Tiny} [36]	0.17B	2023/05/09	1.18	1.84	3.94	6.35
	GroundingDINO _{Swin-Base} [36]	0.23B	2023/05/09	1.97	2.23	6.43	7.58
Specialized LVLMs for REC	Shikra [†] [8]	7B	2023/07/03	0.80	0.52	2.26	1.60
	ONE-PEACE _{Grounding} [‡] [52]	4B	2023/07/20	1.02	0.51	2.64	1.34
	Kosmos-2 [41]	1.6B	2023/10/30	8.06	4.20	8.64	4.52
	CogVLM _{Grounding} [55]	7B	2023/11/20	15.38	8.33	30.73	18.15
	Griffon [64]	13B	2023/12/06	9.16	4.81	16.77	9.18
	Ferret [60]	7B	2023/12/14	3.18	1.62	8.48	4.43
	Rex-Omni [21]	3B	2025/10/15	54.06	37.10	43.90	28.52
General LVLMs 0 ~ 7B	DeepSeek-VL2 _{Tiny} [57]	3B	2024/12/13	2.35	1.84	4.51	4.98
	Qwen2.5-VL [4]	3B	2025/02/20	40.00	25.06	38.22	23.89
	Qwen3-VL [49]	4B	2025/10/15	48.68	32.31	47.78	32.28
General LVLMs 7B ~ 10B	Qwen-VL [3]	7B	2023/08/12	10.91	5.77	18.32	10.08
	MiniGPT-v2 [6]	7B	2023/10/13	2.69	1.36	6.38	3.29
	InternVL2.5 [†] [9]	8B	2024/12/05	0.58	1.14	1.79	4.2
	Qwen2.5-VL [4]	7B	2025/02/20	42.68	27.20	41.46	26.52
	GLM 4.1V [50]	9B	2025/07/01	24.39	14.05	30.57	18.74
	Ovis2.5 [38]	9B	2025/08/19	4.24	2.20	14.56	8.39
	MiMo-VL _{RL} [48]	7B	2025/08/21	19.75	11.22	30.24	18.79
	InternVL3.5 [†] [56]	8B	2025/08/26	8.17	4.48	19.73	11.8
	Qwen3-VL [49]	8B	2025/10/15	51.66	35.13	46.36	31.85
General LVLMs 10B ~ 30B	SPHINX-v2 [†] [29]	13B	2023/11/17	1.59	0.80	4.61	2.36
	DeepSeek-VL2 _{Small} [57]	16B	2024/12/13	34.95	21.34	36.15	22.81
	DeepSeek-VL2 [57]	27B	2024/12/13	25.31	14.88	29.46	19.08
	Qwen3-VL [49]	30B	2025/10/15	56.88	39.98	51.09	35.58
General LVLMs > 30B	GLM 4.5V [50]	106B	2025/08/11	35.95	21.97	40.21	25.41
	Qwen3-VL [†] [49]	235B	2025/10/04	58.79	41.93	52.16	36.89
Closed-Source (APIs) Models	DINO-XSeek [44]	–	2025/03/11	54.47	37.46	46.11	30.14
	Gemini 2.5 Pro [10]	–	2025/03/25	3.45	1.91	8.11	4.90
	Seed1.5-VL [16]	–	2025/05/12	43.52	27.84	37.27	23.03
	Qwen3-VL-Plus [49]	–	2025/09/22	58.11	41.32	50.98	36.18

[†] State-of-the-art method. [‡] Models only predict a single bounding box, limiting multi-target performance.

* GPT and Claude are excluded due to output bounding-box format failures.

Table 3. Results of fine-tuning baselines on RefDrone benchmark.

Methods	F1 _{inst.}	Acc _{inst.}	F1 _{img.}	Acc _{img.}
GLIP-T [27]	56.92	40.39	41.31	28.88
GDINO-T [36]	67.64	51.55	54.54	39.63
NGDINO-T (Ours)	71.11	55.52	56.51	41.20
GDINO-B [36]	69.75	53.95	56.95	41.81
NGDINO-B (Ours)	72.51	57.22	57.84	42.54

ple multi-target structure in gRefCOCO, where expressions predominantly reference only one or two objects. On the more complex aerial-view RSVG [63] benchmark (Table 5), NGDINO-T consistently outperforms prior state-of-the-art methods across all metrics. These results confirm that our proposed modifications effectively enhance performance on diverse multi-target REC tasks.

Table 4. Experimental results on gRefCOCO [30] dataset. Asterisk (*) denotes results reported in the gRefCOCO paper.

Methods	testA		testB	
	Pr@0.5 ↑	N-acc. ↑	Pr@0.5 ↑	N-acc. ↑
MDETR* [22]	50.0	34.5	36.5	31.0
UNINEXT* [59]	46.4	49.3	42.9	48.2
GDINO-T [36]	61.9	83.6	53.7	79.6
NGDINO-T	62.7	83.4	54.0	80.7

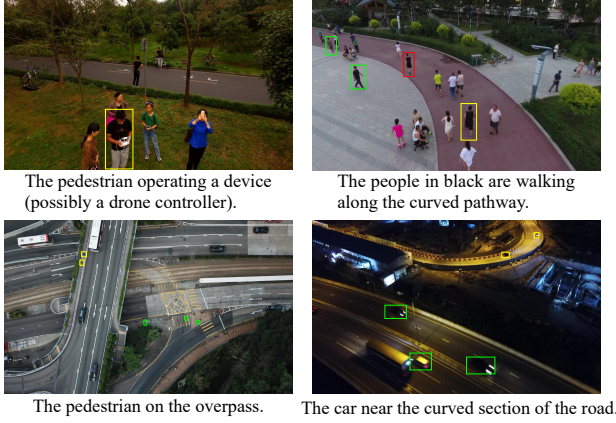


Figure 7. Some failure cases of NGDINO on RefDrone dataset. Red, green, and yellow boxes indicate true positives, false positives, and false negatives, respectively.

5.3. Ablation studies

Analysis of NGDINO components. Table 6 presents the analysis of each component in NGDINO. With only the number prediction head, the model achieves improvements, improving $F1_{inst.}$ from 67.64% to 69.73% (+2.09%). This demonstrates that the auxiliary number prediction task enhances the model’s grounding capability. The number cross-attention mechanism alone yields more modest gains, improving $F1_{inst.}$ to 68.88% (+1.24%). This improvement can be partially attributed to the additional parameters introduced in the decoder. The most significant performance boost occurs when combining both components, where $F1_{inst.}$ reaches 71.11% (+3.47%). This substantial accuracy improvement comes at a negligible cost to efficiency, with only a minor decrease in inference speed (from 13.5 to 12.3 FPS).

Number prediction accuracy. We evaluate the number prediction head’s performance. The head achieves a mean absolute error (MAE) of 0.21, indicating that its numerical predictions are highly precise and closely align with the ground truth counts. This strong performance in number prediction is a key factor in the model’s ability to handle multi-target expressions. The overall accuracy for predicting the correct number of instances is 75.3%.

Table 5. Experimental results on RSVG [63] dataset. Asterisk (*) denotes results reported in the EarthGPT paper. Pr@0.5: percentage of predictions at $IoU \geq 0.5$.

Methods	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
TransVG* [11]	72.41	67.38	60.05	49.10	27.84
MGVLF* [63]	76.78	72.68	66.74	56.42	35.07
EarthGPT* [65]	76.65	71.93	66.52	56.53	37.63
GDINO-T [36]	76.99	75.81	73.36	66.77	49.96
NGDINO-T	77.16	76.03	73.81	67.84	51.39

Table 6. Ablation study of NGDINO components on the RefDrone dataset with Swin-Tiny backbone.

Number Prediction Head	Number Cross-Attention	$F1_{inst.}$	$Acc_{inst.}$	$F1_{img.}$	$Acc_{img.}$	FPS
✓		67.64	51.55	54.54	39.63	13.5
	✓	69.73	53.90	56.29	41.08	12.8
✓	✓	68.88	52.89	54.98	39.74	12.9
		71.11	55.52	56.51	41.20	12.3

5.4. Limitations

Despite its strong performance on multi-target and no-target scenarios, NGDINO has several limitations. As illustrated by the qualitative examples in Figure 7, failure cases typically arise from challenging scenarios inherent to the RefDrone dataset. These failures can be categorized into three primary sources: (1) complexity in the referring expression that demands sophisticated contextual reasoning; (2) cluttered backgrounds that camouflage target objects; and (3) inherent difficulties in detecting very small-scale objects. Addressing these challenging, real-world conditions remains an important direction for future work.

6. Conclusion

In this work, we introduce RefDrone, a challenging benchmark for referring expression comprehension in drone scenes. The dataset is built using RDAnnotator, an efficient semi-automated annotation pipeline that combines LVLMS with human-in-the-loop verification to ensure high-quality annotations. Our extensive experiments reveal a substantial performance drop for existing state-of-the-art methods when evaluated on RefDrone, underscoring the benchmark’s difficulty and exposing key limitations in current REC approaches. Furthermore, we develop NGDINO to address the multi-target and no-target challenges in RefDrone. In the future, we aim to further enhance NGDINO to address additional challenges presented by RefDrone. We also plan to extend RefDrone to larger-scale scenarios and introduce additional tasks such as referring expression segmentation and tracking. We believe RefDrone will serve as a valuable benchmark for advancing research in drone-based REC tasks.

7. Acknowledgment

This work was supported in part by NSFC 62222112, 62225113, 62325111, and 62176186, the Innovative Research Group Project of Hubei Province under Grants (2024AFA017).

References

- [1] Telmo Adão, Jonáš Hruška, Luís Pádua, José Bessa, Emanuel Peres, Raul Morais, and Joaquim Joao Sousa. Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry. *Remote sensing*, 9(11):1110, 2017. 1
- [2] Robert L Axtell. Zipf distribution of us firm sizes. *science*, 293(5536):1818–1820, 2001. 5
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3, 7
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 7, 13
- [5] Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *IEEE Int. Conf. Robot. Autom.*, pages 5228–5234, 2024. 1
- [6] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 3, 7
- [7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 12
- [8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3, 6, 7
- [9] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 7
- [10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Bliestein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7
- [11] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *Int. Conf. Comput. Vis.*, pages 1769–1779, 2021. 8
- [12] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *ACM Int. Conf. Multimedia*, pages 11198–11201, 2024. 12
- [13] Yue Fan, Winsen Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang, and Xin Wang. Aerial vision-and-dialog navigation. In *Findings of ACL*, pages 3043–3061, 2023. 1
- [14] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Adv. Neural Inform. Process. Syst.*, 33:6616–6628, 2020. 3
- [15] Chen Gao, Baining Zhao, Weichen Zhang, Jun Zhang, Jinzhu Mao, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, Xinlei Chen, and Yong Li. EmbodiedCity: A benchmark platform for embodied agent in real-world city environment. *arXiv preprint arXiv:2410.09604*, 2024. 1
- [16] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 7, 13
- [17] Zeyu Han, Fangrui Zhu, Qianru Lao, and Huaizu Jiang. Zero-shot referring expression comprehension via structural similarity between images and captions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14364–14374, 2024. 3
- [18] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(2):684–696, 2019.
- [19] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1115–1124, 2017. 3
- [20] Wei Ji, Li Li, Hao Fei, Xiangyan Liu, Xun Yang, Juncheng Li, and Roger Zimmermann. Towards complex-query referring image segmentation: A novel benchmark. *arXiv preprint arXiv:2309.17205*, 2023. 2, 5
- [21] Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. *arXiv preprint arXiv:2510.12798*, 2025. 3, 6, 7, 13
- [22] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *Int. Conf. Comput. Vis.*, pages 1780–1790, 2021. 3, 6, 7, 8
- [23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proc. EMNLP*, pages 787–798, 2014. 2

- [24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123:32–73, 2017. 2
- [25] Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. CityNav: Language-goal aerial navigation dataset with geographic information. *arXiv preprint arXiv:2406.14240*, 2024. 1
- [26] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 2
- [27] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10965–10975, 2022. 3, 6, 7, 12, 13
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 2
- [29] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. SPHINX: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 3, 6, 7
- [30] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23592–23601, 2023. 2, 4, 6, 8
- [31] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Int. Conf. Comput. Vis.*, pages 4673–4682, 2019. 3
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Adv. Neural Inform. Process. Syst.*, 36, 2023. 3
- [33] Junli Liu, Qizhi Chen, Zhigang Wang, Yiwen Tang, Yiting Zhang, Chi Yan, Dong Wang, Xuelong Li, and Bin Zhao. AerialVLG: A challenging benchmark for aerial visual grounding by exploring positional relations. In *Int. Conf. Comput. Vis.*, pages 5177–5187, 2025. 2
- [34] Kehui Liu, Zixin Tang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. COHERENT: Collaboration of heterogeneous multi-robot system with large language models. In *IEEE Int. Conf. Robot. Autom.*, pages 10208–10214, 2025. 1
- [35] Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. DQ-DETR: Dual query detection transformer for phrase extraction and grounding. In *Proc. AAAI Conf. Artif. Intell.*, volume 37, pages 1728–1736, 2023. 3
- [36] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. *Eur. Conf. Comput. Vis.*, 2024. 3, 5, 6, 7, 8, 12, 13
- [37] Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yan-ning Zhang, and Qi Wu. AerialVLN: Vision-and-language navigation for uavs. In *Int. Conf. Comput. Vis.*, pages 15384–15394, 2023. 1
- [38] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jian-shan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025. 7
- [39] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11–20, 2016. 2
- [40] Kunyu Peng, Jia Fu, Kailun Yang, Di Wen, Yufan Chen, Ruiping Liu, Junwei Zheng, Jiaming Zhang, M. Saquib Sarfraz, Rainer Stiefelhausen, and Alina Roitberg. Referring atomic video action recognition. In *Eur. Conf. Comput. Vis.*, 2024. 2
- [41] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *Int. Conf. Learn. Represent.*, 2024. 3, 7
- [42] Mengxue Qu, Yu Wu, Wu Liu, Xiaodan Liang, Jingkuan Song, Yao Zhao, and Yunchao Wei. RIO: A benchmark for reasoning intention-oriented objects in open environments. *Adv. Neural Inform. Process. Syst.*, 36, 2023. 2
- [43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 13
- [44] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, et al. DINO-X: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024. 6, 7, 13
- [45] Khin Thida San, Sun Ju Mun, Yeong Hun Choe, and Yoon Seok Chang. Uav delivery monitoring system. In *MATEC Web of Conferences*, volume 151, page 04011, 2018. 1
- [46] Qie Sima, Sinan Tan, Huaping Liu, Fuchun Sun, Weifeng Xu, and Ling Fu. Embodied referring expression for manipulation question answering in interactive environment. In *IEEE Int. Conf. Robot. Autom.*, 2023. 1
- [47] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. ReCLIP: A strong zero-shot baseline for referring expression comprehension. In *Proc. ACL*, 2022. 3, 13, 14
- [48] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, et al. MiMo-VL Technical Report. *arXiv preprint arXiv:2506.03569*, 2025. 7, 13
- [49] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6, 7, 13
- [50] V Team. GLM-4.5V and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 7, 13
- [51] Leon Amadeus Varga, Benjamin Kiefer, Martin Messmer, and Andreas Zell. SeaDronesSee: A maritime benchmark for detecting humans in open water. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2260–2270, 2022. 1

- [52] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. ONE-PEACE: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023. [6](#), [7](#)
- [53] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. EmbodiedScan: A holistic multi-modal 3d perception suite towards embodied ai. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19757–19767, 2024. [1](#)
- [54] Tianyu Wang, Haitao Lin, Junqiu Yu, and Yanwei Fu. Polaris: Open-ended interactive robotic manipulation via syn2real visual grounding and large language models. In *International Conference on Intelligent Robots and Systems*, 2024. [1](#)
- [55] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. CogVLM: Visual expert for pretrained language models. *Adv. Neural Inform. Process. Syst.*, 37:121475–121499, 2024. [3](#), [7](#)
- [56] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. [7](#)
- [57] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, , et al. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. [3](#), [6](#), [7](#), [13](#)
- [58] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. *Adv. Neural Inform. Process. Syst.*, 36, 2023. [2](#)
- [59] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15325–15336, 2023. [3](#), [8](#)
- [60] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *Int. Conf. Learn. Represent.*, 2024. [3](#), [7](#)
- [61] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Eur. Conf. Comput. Vis.*, pages 69–85, 2016. [1](#), [2](#), [4](#)
- [62] Sergey Zagoruyko. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. [3](#), [12](#)
- [63] Yang Zhan, Zhitong Xiong, and Yuan Yuan. RSVG: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. [2](#), [5](#), [6](#), [7](#), [8](#)
- [64] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *Eur. Conf. Comput. Vis.*, pages 405–422, 2024. [3](#), [7](#)
- [65] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. EarthGPT: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024. [8](#)
- [66] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3DRefer: Grounding text description to multiple 3d objects. In *Int. Conf. Comput. Vis.*, pages 15225–15236, 2023. [5](#)
- [67] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7380–7399, 2021. [2](#), [3](#), [4](#)