# GeoBox-R1: Curriculum-Guided SFT and Geometric RL for Unified Box-Level Remote Sensing Visual Grounding

**Anonymous submission**


## Abstract

Remote sensing visual grounding (RSVG) aims to localize the object referred to by a natural-language expression in aerial or satellite imagery. Recent multimodal large language models (MLLMs) make unified box-level grounding feasible, but general-purpose models do not reliably produce oriented bounding boxes (OBBs) and often fall back to horizontal bounding box (HBB) outputs. Jointly training a single model for HBB and OBB prediction is also challenging because the two tasks require different levels of semantic grounding and geometric precision. To address this issue, we propose **GeoBox-R1**, a two-stage training paradigm for unified box-level RSVG. It first performs curriculum-guided supervised fine-tuning (SFT), ordering training data from HBB grounding to OBB grounding and finally HBB-to-OBB Chain-of-Thought supervision to establish unified box-level grounding progressively. It then applies geometric reinforcement learning via Group Reward-Decoupled Normalization Policy Optimization (GDPO), using two rule-based geometric rewards, Rotated Intersection over Union (RIoU) and adaptive Wasserstein distance, to refine OBB prediction. Experiments on 7 HBB and 3 OBB benchmarks show that GeoBox-R1 achieves the best macro averages on all three metrics, Acc@0.5, Acc@0.7, and mean overlap, for both tasks. With only 4B parameters, it surpasses state-of-the-art 7B–8B models, leading by 3.68/4.89/2.61 points on HBB and by 5.36/7.74/2.89 points on OBB, with the largest gains at the stricter Acc@0.7 threshold.


## Introduction

Remote sensing visual grounding (RSVG), which localizes objects in overhead imagery according to natural-language descriptions, is relevant to applications such as maritime ship monitoring (Baek et al. 2024), aircraft analysis (El Ghazouali et al. 2024), and urban observation (Gong, Liu, and Huang 2020). Early RSVG methods predominantly adopt horizontal bounding boxes (HBBs). However, HBBs cannot encode target orientation or tightly fit rotated objects, whereas oriented bounding boxes (OBBs) provide a more accurate and compact representation but require estimating orientation and corner locations under limited annotations. Rather than replacing HBB with OBB entirely, it is therefore more practical to preserve the well-established HBB grounding capability while extending the model to OBB prediction—a setting we refer to as *unified box-level grounding*.

Recent multimodal large language models (MLLMs) make this unified setting increasingly feasible. Large-scale multimodal pretraining provides strong vision-language alignment and instruction following for complex referring expressions in remote sensing scenes (Liu et al. 2024; Wang et al. 2024; Bai et al. 2025b,a), and spatial coordinates can be represented as text tokens (Chen et al. 2023; Peng et al. 2023), so grounding proceeds autoregressively without task-specific detection heads, flexibly accommodating different box representations in one generation pipeline.

However, directly extending MLLMs to OBB is nontrivial. As Fig. 1 shows, even a strong recent general-purpose MLLM such as Qwen3-VL still returns an orientation-agnostic horizontal box when explicitly prompted for an OBB, showing that prompting alone does not reliably elicit oriented grounding. Supervised fine-tuning can teach this output behavior, but two obstacles remain. For capability acquisition, HBB supervision drives coarse semantic localization while OBB supervision demands precise geometry over corners, aspect ratio, and orientation; mixing these heterogeneous signals can blur the coarse-to-fine trajectory, and Chain-of-Thought (CoT) reasoning (Wei et al. 2022)—predicting an HBB anchor and refining it into an OBB—bridges the two levels at the cost of an extra intermediate structure to learn. For localization precision, token-level SFT treats coordinates as ordinary text and does not directly optimize continuous geometric quality such as overlap or orientation accuracy, so OBB predictions may remain coarse even when correctly formatted.

To address these challenges, we propose **GeoBox-R1**, a two-stage training paradigm for unified box-level RSVG. Motivated by curriculum learning (Bengio et al. 2009), the first stage performs *curriculum-guided supervised fine-tuning (SFT)*: within a single run, training proceeds easy-to-hard from HBB grounding to OBB grounding and finally HBB-to-OBB CoT supervision, so the model adapts progressively from coarse semantic grounding to fine-grained geometric reasoning, with CoT providing process supervision that bridges the two. The second stage addresses the

This is an anonymized submission for review purposes only. Distribution, citation, or public sharing of this manuscript is strictly prohibited. Copyright and publication details will appear in the final version if accepted.

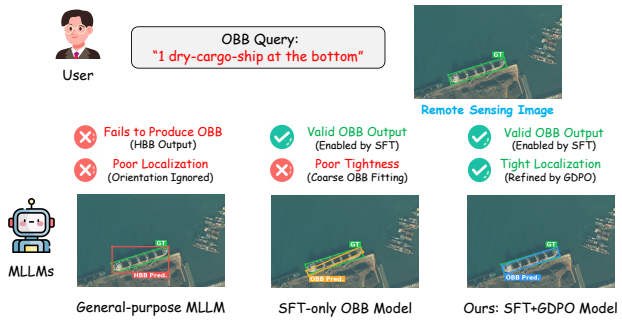



Figure 1: Motivation for unified box-level RSVG. Given an OBB query, a general-purpose MLLM can fail to produce an OBB and instead return an HBB that ignores target orientation. SFT enables valid OBB outputs, yet the SFT-only model exhibits coarse fitting; subsequent geometric RL (GDPO) produces a tighter oriented box.

remaining precision gap with geometric reinforcement learning (RL) based on Group Reward-Decoupled Normalization Policy Optimization (GDPO) (Liu et al. 2026), using Rotated Intersection over Union (RIoU) (Yang et al. 2023), aligned with the evaluation metric, and adaptive Wasserstein distance (WD) for dense geometric feedback in low-overlap and cross-scale cases.

Our contributions are summarized as follows:

- We propose GeoBox-R1, a two-stage training paradigm that treats unified box-level RSVG as coarse-to-fine capability acquisition.
- We introduce a curriculum-guided SFT scheme that sequences HBB, OBB, and HBB-to-OBB CoT supervision within a single run.
- We propose a geometry-aware RL strategy for OBB refinement, combining RIoU and adaptive WD rewards under GDPO.
- On 7 HBB and 3 OBB benchmarks under the unified refGeo (Zhou et al. 2024) setting, our 4B model achieves the best macro averages on Acc@0.5, Acc@0.7, and mean overlap for both tasks, surpassing state-of-the-art 7B–8B models by 3.68/4.89/2.61 points on HBB and 5.36/7.74/2.89 points on OBB.

## Related Work

### Remote Sensing Visual Grounding

GeoVG (Sun et al. 2022) first formalized RSVG and showed that natural-image grounding methods degrade sharply under the scale variation and clutter of overhead scenes. Zhan et al. (Zhan, Xiong, and Yuan 2023) built the large-scale DIOR-RSVG benchmark with multi-granularity visual-language fusion for small targets. Later works adapted TransVG (Deng et al. 2021) and QRNet (Ye et al. 2022) to remote sensing and introduced language-guided progressive attention (Li et al. 2024). These methods use separate unimodal encoders and task-specific fusion heads, which can limit generalization to complex referring expressions.

MLLMs introduce a language-generation paradigm for RSVG. GeoChat (Kuckreja et al. 2024) and SkyEyeGPT (Zhan, Xiong, and Yuan 2025) formulate remote sensing grounding within unified language generation, GeoPixel (Shabbir et al. 2025) extends grounding to pixel-level mask prediction, GeoGround (Zhou et al. 2024) unifies HBB, OBB, and mask grounding in one remote-sensing MLLM, and RSGround-R1 (Huang, He, and Wen 2026) couples CoT SFT with spatially guided Group Relative Policy Optimization (GRPO). VRSBench (Li, Ding, and Elhoseiny 2024) and AerialVG (Liu et al. 2025a) further stress relational and positional reasoning. These efforts advance architectures and data, but the training paradigm—how to organize SFT for a stable HBB-to-OBB transition and how to design RL rewards for oriented geometry—remains underexplored.

### MLLM Post-Training for Visual Grounding

**SFT Strategies.** Early works such as Shikra (Chen et al. 2023) and Kosmos-2 (Peng et al. 2023) establish that spatial coordinates can be represented as text tokens via referential dialogue and grounded text generation. Building on this, grounding-oriented MLLMs such as GROUNDHOG (Zhang et al. 2024), GLaMM (Rasheed et al. 2024), LISA (Lai et al. 2024), F-LMM (Wu et al. 2025), and Ground-V (Zong et al. 2025) extend grounding to segmentation, panoptic masks, and multi-step instructions. These models support multiple grounding outputs, but they generally do not study how to order HBB and OBB supervision when both are learned by one autoregressive model. Directly mixing coarse localization and fine geometric supervision can make this joint setting harder to optimize. In contrast, our curriculum-guided SFT sequences HBB, OBB, and CoT supervision in an easy-to-hard order, consolidating coarse localization before fine geometric prediction.

**RL and Reward Design.** DeepSeekMath (Shao et al. 2024) introduces GRPO, replacing Proximal Policy Optimization (PPO)'s value model with groupwise advantage estimation, and DeepSeek-R1-Zero (Guo et al. 2025) shows that GRPO with rule-based accuracy and format rewards can elicit reasoning without supervised reasoning traces. Visual-RFT (Liu et al. 2025b) designs verifiable visual rewards such as detection Intersection over Union (IoU), while R1-Onevision (Yang et al. 2025) and R1-VL (Zhang et al. 2025) refine multimodal reasoning with step- and process-level group optimization. For coexisting rewards, GDPO (Liu et al. 2026) decouples group-wise normalization per reward channel and stabilizes optimization via batch-wise advantage normalization. These reward formulations, however, target reasoning correctness or semantic quality; the geometric properties central to OBBs—corner alignment, orientation, and aspect-ratio fidelity—call for dedicated reward signals.

## Methodology

### Unified Box-Level Formulation

Given a remote sensing image $I$ and a referring expression $q$, unified box-level RSVG localizes the referred target with

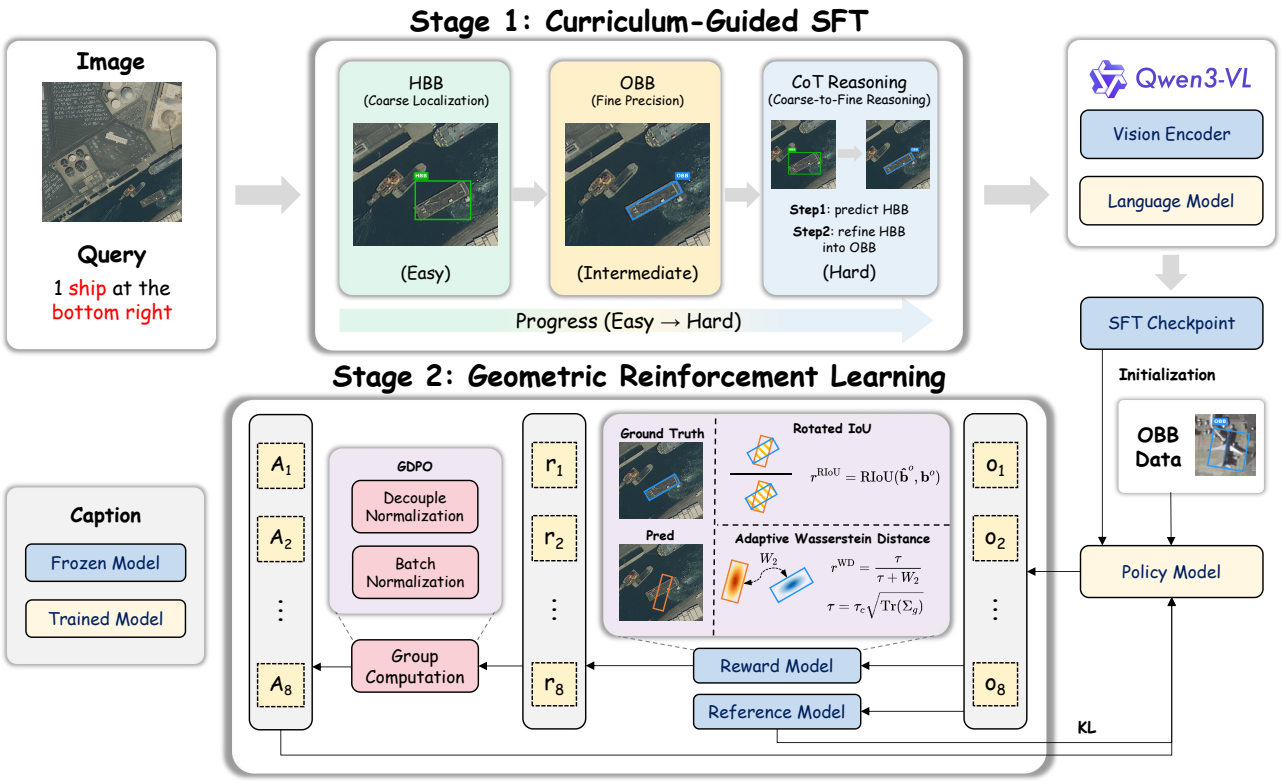



Figure 2: Overview of GeoBox-R1. In Stage 1, curriculum-guided SFT organizes HBB grounding, OBB grounding, and HBB-to-OBB CoT reasoning in an easy-to-hard sequence within a single SFT run. In Stage 2, OBB-only geometric RL samples multiple completions per prompt and optimizes the model with GDPO using RIoU and adaptive WD rewards to improve oriented localization accuracy.

a single autoregressive MLLM under a unified question-answer interface. Rather than introducing an explicit task token, we encode the requested output format directly in the question template:

$$x = \phi_m(q), \quad m \in \{\mathrm{HBB}, \mathrm{OBB}, \mathrm{CoT}\}, \tag{1}$$

where $\phi_{\mathrm{HBB}}(\cdot)$ requests a horizontal box, $\phi_{\mathrm{OBB}}(\cdot)$ an oriented box, and $\phi_{\mathrm{CoT}}(\cdot)$ stepwise localization. The index $m$ is not passed as an explicit task token; the requested format is conveyed through the instantiated question $x$. Each training sample is thus a triple $(I, x, y)$ with target answer $y$.

Following Qwen3-VL (Bai et al. 2025a), all coordinates are quantized into the range $[0, 1000]$. For horizontal grounding, the target is represented as an HBB $\mathbf{b}^h = [x_1, y_1, x_2, y_2]$, where $(x_1, y_1)$ and $(x_2, y_2)$ denote the top-left and bottom-right corners; for oriented grounding, the target is represented as an OBB $\mathbf{b}^o = [\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4]$ with $\mathbf{p}_i = [x_i, y_i]$, whose four corners are listed in the storage order of the corresponding dataset. For $\phi_{\mathrm{CoT}}$, the target response contains an intermediate HBB for coarse localization and a final OBB for refined localization, making the coarse-to-fine transition explicit in the output space. All grounding variants are then cast as conditional sequence generation:

$$p_\theta(y \mid I, x), \tag{2}$$

where the same parameters $\theta$ are shared across HBB grounding, OBB grounding, and HBB-to-OBB reasoning traces. GeoBox-R1 optimizes this unified interface in two stages: curriculum-guided SFT first establishes coarse-to-fine grounding capability, and geometric RL subsequently improves OBB precision. The overall framework is shown in Fig. 2.

## Curriculum-Guided SFT

**Curriculum Design.** Although all samples target the same grounding task, the supervision signals induced by different templates vary in difficulty and focus. HBB supervision mainly teaches coarse semantic localization under a compact schema; OBB supervision further requires precise orientation, aspect ratio, and corner prediction; CoT supervision is the most demanding, additionally requiring an explicit coarse-to-fine refinement trace. Directly mixing them forces the model to acquire semantic grounding, geometric precision, and structured refinement simultaneously. We therefore partition the SFT data and arrange it in a single-pass easy-to-hard curriculum:

$$\mathcal{D}_{\mathrm{SFT}} = \mathcal{D}_{\mathrm{HBB}} \oplus \mathcal{D}_{\mathrm{OBB}} \oplus \mathcal{D}_{\mathrm{CoT}}, \tag{3}$$

where $\oplus$ denotes sequential concatenation: HBB samples establish coarse localization, OBB samples introduce oriented

geometry, and CoT samples provide explicit HBB-to-OBB refinement supervision. Instead of a stage-wise pipeline with checkpoint handoffs, we adopt a single continuous SFT run with sequential data ordering, which preserves the easy-to-hard progression.

**Response Format.** The three subsets use the same image-expression input schema but differ in their data sources and target response formats. Responses are serialized as JSON: HBB samples answer with `{"horizontal_bbox": [x1,y1,x2,y2]}` and OBB samples with `{"oriented_bbox": [[x1,y1],...,[x4,y4]]}`, respectively providing compact horizontal-localization supervision and more detailed geometric supervision over orientation, aspect ratio, and corner coordinates. For samples in $\mathcal{D}_{\mathrm{CoT}}$, the response is a structured Chain-of-Thought that records the HBB-to-OBB refinement trace $y_{\mathrm{CoT}} = \mathbf{b}^h \rightarrow \mathbf{b}^o$; unlike the single-step HBB or OBB responses, it supervises both the intermediate refinement step and the final prediction, yielding process-level supervision. Each CoT target is synthesized from a dual-annotated sample: the intermediate $\mathbf{b}^h$ is the horizontal annotation (the axis-aligned envelope of the OBB corners) and the final $\mathbf{b}^o$ is the oriented annotation, rendered into a fixed two-step template. The trace is programmatic rather than a free-form rationale: it decomposes oriented grounding into a locate-then-refine procedure with a supervised intermediate step. The ablation on data organization and CoT (Table 3) supports using this structured supervision within the proposed curriculum.

**SFT Objective.** Let $\mathcal{D}_{\mathrm{SFT}} = \{(I_n, x_n, y_n)\}_{n=1}^{N}$ denote the full first-stage training set. We train with the standard autoregressive negative log-likelihood:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{n=1}^{N} \log p_\theta(y_n \mid I_n, x_n). \tag{4}$$

Because samples follow the curriculum order of Eq. (3), the same token-level objective induces coarse-to-fine capability acquisition without extra losses or task-specific heads.

## Geometric RL with GDPO

After curriculum-guided SFT, we initialize RL from that checkpoint and optimize on OBB data only: the goal is not to relearn coarse grounding but to sharpen geometric precision. We adopt GDPO (Liu et al. 2026), which performs reward-decoupled group-wise normalization across reward channels followed by batch-wise advantage normalization. Given an image-prompt pair $(I, x)$, the RL stage samples a group of $G$ completions $\{y_i\}_{i=1}^{G}$ from the current policy and performs one policy update per group. Each completion is parsed into an OBB JSON output; a completion that cannot be parsed is retained in the group but receives zero reward on all channels, without an additional format-validity reward.

**RIoU Reward.** The first reward directly measures oriented localization quality, $r^{\mathrm{RIoU}} = \mathrm{RIoU}(\hat{\mathbf{b}}^o, \mathbf{b}^o)$, where $\hat{\mathbf{b}}^o$ and $\mathbf{b}^o$ are the predicted and ground-truth OBBs. It is directly aligned with the OBB evaluation protocol but becomes sparse when the overlap is zero.

**Adaptive Wasserstein Distance Reward.** To complement the sparsity of RIoU, we introduce a continuous geometry-aware reward based on the Wasserstein distance, which supplies denser feedback in low-overlap and large-scale-variation cases. Inspired by Gaussian-based formulations for rotated object modeling (Yang et al. 2021, 2023; Hou et al. 2023), we convert each OBB into a 2D Gaussian before computing WD: the four corners $\{\mathbf{p}_i\}_{i=1}^{4}$ are regarded as observations from a Gaussian distribution whose parameters are estimated by maximum likelihood, yielding the closed-form mean and covariance

$$\boldsymbol{\mu} = \frac{1}{4}\sum_{i=1}^{4} \mathbf{p}_i, \quad \boldsymbol{\Sigma} = \frac{1}{4}\sum_{i=1}^{4} (\mathbf{p}_i - \boldsymbol{\mu})(\mathbf{p}_i - \boldsymbol{\mu})^\top. \tag{5}$$

This formulation matches the autoregressive OBB output format directly, avoids an extra angle canonicalization step, and lets the covariance capture orientation and spatial extent. The squared Wasserstein distance between the predicted Gaussian $(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and the ground-truth Gaussian $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is

$$W_2^2 = \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_g\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_g - 2\left(\boldsymbol{\Sigma}_g^{1/2}\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_g^{1/2}\right)^{1/2}\right), \tag{6}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace, and the distance is $W_2 = \sqrt{W_2^2}$. We convert it into a bounded reward with an adaptive scale:

$$r^{\mathrm{WD}} = \frac{\tau}{\tau + W_2}, \quad \tau = \tau_c \sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_g)}, \tag{7}$$

where $\tau_c$ is a scaling factor. For the same relative localization error, $W_2$ increases with object size, so a fixed $\tau$ assigns systematically lower rewards to larger targets. Scaling $\tau$ with $\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma}_g)}$ normalizes reward sensitivity with respect to target size and makes the feedback more comparable across scales.

**GDPO Multi-Reward Aggregation.** The two reward channels, RIoU and adaptive WD, are equally weighted. Following GDPO (Liu et al. 2026), we first perform per-reward group-wise normalization within each sampled group, aggregate the normalized rewards, and then apply batch-wise advantage normalization before the policy update. For reward type $k$, the group statistics over the $G$ sampled completions are

$$\bar{r}^{(k)} = \frac{1}{G}\sum_{j=1}^{G} r_j^{(k)}, \quad \sigma^{(k)} = \sqrt{\frac{1}{G}\sum_{j=1}^{G}\left(r_j^{(k)} - \bar{r}^{(k)}\right)^2}. \tag{8}$$

The groupwise aggregated advantage for completion $i$ is

$$\tilde{A}_i = \sum_{k \in \{\mathrm{RIoU}, \mathrm{WD}\}} \lambda_k \frac{r_i^{(k)} - \bar{r}^{(k)}}{\sigma^{(k)} + \varepsilon}, \tag{9}$$

followed by batch-wise advantage normalization before the policy update:

$$\hat{A}_i = \frac{\tilde{A}_i - \mu_{\mathcal{B}}(\tilde{A})}{\sigma_{\mathcal{B}}(\tilde{A}) + \varepsilon}, \tag{10}$$

where $\mu_{\mathcal{B}}(\tilde{A})$ and $\sigma_{\mathcal{B}}(\tilde{A})$ are computed over the RL mini-batch $\mathcal{B}$ and $\varepsilon$ is a small constant. Reward-decoupled aggregation prevents channel-scale differences from dominating the combined advantage while retaining within-channel

Table 1: HBB results. Each benchmark reports Acc@0.5/0.7 and mIoU (%); HBB Avg. is the macro average. Best/second-best results are shown in **bold**/underline. * indicates modified test sets; all baselines are re-evaluated by us with model-specific grounding prompts and output parsers. Conventions apply to later tables.

| Model | Params | DIOR-RSVG-Test | | | DIOR-RSVG-Val | | | RSVG-Test | | | RSVG-Val | | | GeoChat* | | | VRSBench* | | | AVVG | | | HBB Avg. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | @0.5 | @0.7 | mIoU | @0.5 | @0.7 | mIoU | @0.5 | @0.7 | mIoU | @0.5 | @0.7 | mIoU | @0.5 | @0.7 | mIoU | @0.5 | @0.7 | mIoU | @0.5 | @0.7 | mIoU | @0.5 | @0.7 | mIoU |
| *General-Purpose MLLMs* | | | | | | | | | | | | | | | | | | | | | | | | | |
| InternVL3 (Zhu et al. 2025) | 2B | 15.01 | 6.95 | 20.74 | 13.60 | 6.61 | 19.52 | 0.24 | 0.00 | 1.73 | 0.33 | 0.08 | 2.16 | 11.09 | 5.09 | 14.93 | 5.51 | 1.67 | 11.08 | 2.15 | 0.42 | 3.58 | 6.85 | 2.98 | 10.53 |
| LLaVA-OV-1.5 (An et al. 2025) | 4B | 18.39 | 4.10 | 24.63 | 16.68 | 3.72 | 23.88 | 1.79 | 0.24 | 4.88 | 1.25 | 0.08 | 5.26 | 11.92 | 2.03 | 20.34 | 7.98 | 1.18 | 16.70 | 1.69 | 0.07 | 5.30 | 8.53 | 1.63 | 14.43 |
| LLaVA-OV-1.5 (An et al. 2025) | 8B | 23.59 | 6.36 | 27.58 | 22.63 | 6.14 | 26.91 | 1.14 | 0.16 | 4.08 | 1.50 | 0.25 | 3.86 | 18.42 | 3.33 | 24.19 | 13.39 | 2.55 | 21.44 | 1.65 | 0.10 | 4.02 | 11.76 | 2.70 | 16.01 |
| InternVL3 (Zhu et al. 2025) | 8B | 29.94 | 16.50 | 30.56 | 27.44 | 14.82 | 29.47 | 1.14 | 0.41 | 3.99 | 1.58 | 0.33 | 5.14 | 24.79 | 10.11 | 26.38 | 19.45 | 6.89 | 23.67 | 6.12 | 2.87 | 6.94 | 15.78 | 7.42 | 18.02 |
| Qwen2.5-VL (Bai et al. 2025b) | 3B | 37.52 | 25.98 | 37.47 | 37.90 | 26.72 | 37.29 | 13.69 | 8.15 | 14.62 | 14.32 | 8.91 | 15.48 | 36.07 | 19.68 | 34.83 | 36.74 | 19.10 | 35.25 | 18.83 | 12.56 | 15.36 | 27.87 | 17.30 | 27.18 |
| Qwen2.5-VL (Bai et al. 2025b) | 7B | 40.31 | 26.77 | 39.29 | 38.59 | 25.13 | 38.09 | 14.83 | 7.91 | 15.50 | 16.74 | 10.16 | 16.97 | 41.84 | 21.65 | 38.51 | 41.51 | 21.18 | 38.76 | 21.91 | 15.07 | 17.78 | 30.82 | 18.27 | 29.27 |
| Qwen3-VL (Bai et al. 2025a) | 4B | 51.91 | 38.16 | 48.21 | 51.38 | 37.05 | 47.58 | 33.17 | 20.05 | 30.72 | 32.89 | 19.57 | 30.45 | 49.17 | 28.29 | 43.37 | 51.42 | 28.80 | 45.21 | 23.93 | 18.99 | 19.79 | 42.01 | 27.27 | 37.90 |
| Qwen3-VL (Bai et al. 2025a) | 8B | 54.14 | 41.50 | 50.81 | 53.68 | 40.50 | 49.63 | 33.82 | 20.13 | 31.81 | 34.80 | 19.98 | 30.78 | 50.73 | 29.23 | 44.37 | 53.55 | 29.80 | 46.72 | 25.22 | 18.55 | 20.21 | 43.70 | 28.53 | 39.19 |
| *Remote Sensing MLLMs* | | | | | | | | | | | | | | | | | | | | | | | | | |
| LHRS-Bot (Muhtar et al. 2024) | 7B | 13.15 | 3.02 | 18.37 | 12.56 | 3.24 | 18.22 | 0.65 | 0.08 | 1.68 | 0.33 | 0.08 | 1.04 | 2.16 | 0.30 | 7.24 | 1.21 | 0.22 | 6.14 | 0.15 | 0.02 | 1.31 | 4.32 | 0.99 | 7.71 |
| LHRS-Bot-Nova (Li et al. 2025) | 8B | 31.31 | 12.36 | 30.99 | 29.75 | 11.66 | 29.89 | 2.12 | 0.24 | 4.80 | 1.75 | 0.17 | 4.12 | 26.13 | 8.68 | 26.92 | 12.54 | 2.67 | 19.55 | 1.04 | 0.00 | 3.51 | 14.95 | 5.11 | 17.11 |
| GeoChat (Kuckreja et al. 2024) | 7B | 31.04 | 8.72 | 32.52 | 29.16 | 8.18 | 31.54 | 2.44 | 0.24 | 7.99 | 4.16 | 0.42 | 9.74 | 36.69 | 13.57 | 36.17 | 14.38 | 3.16 | 23.66 | 0.31 | 0.00 | 1.86 | 16.88 | 4.90 | 20.50 |
| *Supervised Fine-Tuning on refGeo* | | | | | | | | | | | | | | | | | | | | | | | | | |
| GeoChat (SFT) (Kuckreja et al. 2024) | 7B | 45.42 | 21.40 | 41.20 | 43.37 | 20.42 | 39.93 | 2.20 | 0.41 | 4.93 | 3.41 | 0.75 | 6.10 | 35.81 | 14.44 | 35.35 | 27.56 | 8.45 | 31.69 | 0.78 | 0.10 | 3.07 | 22.65 | 9.42 | 23.18 |
| InternVL3 (SFT) (Zhu et al. 2025) | 2B | 57.15 | 38.17 | 50.18 | 56.12 | 36.73 | 49.41 | 8.56 | 2.12 | 12.67 | 10.32 | 1.92 | 14.91 | 40.15 | 17.07 | 37.34 | 36.29 | 14.94 | 36.52 | 9.58 | 3.71 | 10.70 | 31.17 | 16.38 | 30.25 |
| LLaVA-OV-1.5 (SFT) (An et al. 2025) | 4B | 61.49 | 40.97 | 53.04 | 59.73 | 39.63 | 51.90 | 20.70 | 5.05 | 24.28 | 22.81 | 6.24 | 24.37 | 45.49 | 16.50 | 40.63 | 37.60 | 14.51 | 38.13 | 10.00 | 2.41 | 11.64 | 36.83 | 17.90 | 34.86 |
| Qwen2.5-VL (SFT) (Bai et al. 2025b) | 3B | 68.09 | 53.42 | 59.85 | 66.72 | 53.23 | 59.09 | 39.85 | 20.54 | 34.68 | 34.55 | 17.32 | 30.74 | 52.44 | 29.47 | 45.66 | 57.26 | 34.31 | 49.19 | 27.11 | 19.73 | 22.36 | 49.43 | 32.57 | 43.08 |
| GeoGround (Zhou et al. 2024) | 7B | **77.70** | **64.98** | **68.28** | **77.13** | 63.00 | **67.54** | 26.65 | 10.02 | 28.15 | 27.81 | 9.08 | 27.08 | **69.76** | **50.66** | **57.97** | 65.76 | 39.05 | **56.27** | 21.60 | 8.37 | 22.28 | 52.35 | 35.02 | 46.80 |
| InternVL3 (SFT) (Zhu et al. 2025) | 8B | 74.83 | 62.76 | 65.99 | 75.06 | 61.57 | 65.71 | 46.05 | 25.51 | 40.08 | 44.05 | 25.81 | 39.19 | 59.23 | 32.03 | 48.16 | 58.05 | 33.70 | 50.57 | 28.41 | 19.91 | 24.74 | 55.10 | 37.33 | 47.78 |
| GeoBox-R1 (SFT) | 4B | 74.91 | 62.46 | 65.99 | 74.21 | 62.23 | 65.55 | 48.57 | 29.42 | 40.51 | 46.29 | 28.73 | 39.75 | 61.39 | 35.13 | 50.34 | 64.45 | 39.58 | 54.28 | 30.33 | 24.74 | 25.32 | 57.17 | 40.33 | 48.82 |
| GeoBox-R1 (SFT + GDPO) | 4B | 76.61 | 64.78 | 67.50 | 75.11 | **63.13** | 66.57 | **51.26** | **32.60** | **42.99** | **48.38** | **31.97** | **42.00** | 61.13 | 35.49 | 51.25 | **66.84** | **41.16** | 55.74 | **32.14** | **26.39** | **26.68** | **58.78** | **42.22** | **50.39** |

preferences; batch-wise normalization then stabilizes the advantage scale under the two geometric objectives.

**Policy Optimization Objective.** With the policy ratio $\rho_i(\theta) = \pi_\theta(y_i \mid I, x)/\pi_{\theta_{\text{old}}}(y_i \mid I, x)$, the policy loss follows the standard clipped group-relative objective:

$$\mathcal{L}_{\text{RL}} = -\frac{1}{G}\sum_{i=1}^{G} \min\Big(\rho_i(\theta)\hat{A}_i, \text{clip}\big(\rho_i(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_i\Big) + \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\,, \tag{11}$$

where $\epsilon$ is the clipping threshold, $\beta$ controls Kullback–Leibler (KL) regularization, and $\pi_{\text{ref}}$ is the frozen curriculum-guided SFT checkpoint. Since this optimization is applied only to OBB samples, it directly targets fine-grained oriented localization.

# Experiments

## Experimental Setup

**Datasets.** We adopt refGeo (Zhou et al. 2024), which unifies RSVG (Sun et al. 2022), DIOR-RSVG (Zhan, Xiong, and Yuan 2023), GeoChat (Kuckreja et al. 2024), VRSBench (Li, Ding, and Elhoseiny 2024), and the newly proposed AVVG. Following the original protocol, we use the filtered GeoChat* and VRSBench* with training samples overlapping DIOR-RSVG val/test images removed to avoid leakage. HBB evaluation covers seven benchmarks (DIOR-RSVG test/val, RSVG test/val, GeoChat*, VRSBench*, AVVG) and OBB evaluation covers the last three. For curriculum-guided SFT, $\mathcal{D}_{\text{HBB}}$ comes from the HBB-only training splits of RSVG and DIOR-RSVG; for the dual-annotated GeoChat, VRSBench, and AVVG, 75% of training samples form the OBB-only $\mathcal{D}_{\text{OBB}}$ and the remaining 25% are converted into HBB-to-OBB CoT supervision as $\mathcal{D}_{\text{CoT}}$. The RL stage uses a random 20% of $\mathcal{D}_{\text{OBB}}$.

**Evaluation Metrics.** Following prior work (Zhou et al. 2024; Huang, He, and Wen 2026; Zhang et al. 2026), we report Acc@0.5, Acc@0.7, and mean overlap (IoU/mIoU for HBB, RIoU/mRIoU for OBB); a prediction is correct at Acc@$\tau$ if its overlap with the ground truth exceeds $\tau$. The unweighted average (Avg.) across evaluation sets is the main aggregate metric.

**Implementation Details.** We instantiate GeoBox-R1 on Qwen3-VL-4B-Instruct (Bai et al. 2025a); both stages use ms-swift (Zhao et al. 2025). The SFT stage adopts LoRA fine-tuning (Hu et al. 2022) with the vision encoder and merger frozen, one epoch, learning rate $1 \times 10^{-4}$, LoRA rank 16, and alpha 32. The RL stage continues from the SFT checkpoint with a frozen copy as $\pi_{\text{ref}}$, using one prompt per group with $G = 8$ rollouts, $\text{num_iterations} = 1$, $\beta = 0.02$, $\tau_c = 8$, $\lambda_{\text{RIoU}} = \lambda_{\text{WD}} = 0.5$, learning rate $5 \times 10^{-6}$, and one epoch. SFT runs on two 24GB RTX 4090 GPUs; RL uses three 40GB A100 GPUs (one vLLM rollout server, two GDPO workers). **GeoBox-R1 (SFT)** denotes the curriculum-guided SFT checkpoint with CoT supervision; **GeoBox-R1 (SFT + GDPO)** applies GDPO with RIoU and adaptive WD rewards on top of it.

## Comparison with State-of-the-Art

**HBB Results.** Table 1 shows that GeoBox-R1 obtains the best HBB macro averages: **58.78** Acc@0.5, **42.22** Acc@0.7,

Table 2: OBB results. Metrics and notation follow Table 1, with mRIoU replacing mIoU; OBB Avg. is the macro average.

| Model | Params | GeoChat* | | | VRSBench* | | | AVVG | | | OBB Avg. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | @0.5 | @0.7 | mRIoU | @0.5 | @0.7 | mRIoU | @0.5 | @0.7 | mRIoU | @0.5 | @0.7 | mRIoU |
| GeoChat (Kuckreja et al. 2024) | 7B | 31.73 | 8.27 | 31.25 | 11.76 | 2.04 | 21.26 | 0.10 | 0.00 | 1.08 | 14.53 | 3.43 | 17.87 |
| InternVL3 (SFT) (Zhu et al. 2025) | 2B | 24.57 | 5.26 | 27.28 | 19.82 | 7.48 | 24.84 | 0.98 | 0.15 | 4.35 | 15.12 | 4.30 | 18.82 |
| GeoChat (SFT) (Kuckreja et al. 2024) | 7B | 28.95 | 11.07 | 30.14 | 21.78 | 5.80 | 27.70 | 0.26 | 0.03 | 1.97 | 17.00 | 5.64 | 19.94 |
| LLaVA-OV-1.5 (SFT) (An et al. 2025) | 4B | 33.23 | 9.76 | 33.18 | 23.87 | 8.71 | 28.54 | 2.69 | 0.59 | 6.14 | 19.93 | 6.35 | 22.62 |
| Qwen2.5-VL (SFT) (Bai et al. 2025b) | 3B | 49.36 | 22.65 | 40.46 | 41.67 | 21.55 | 39.98 | 16.13 | 8.63 | 15.12 | 35.72 | 17.61 | 31.86 |
| InternVL3 (SFT) (Zhu et al. 2025) | 8B | 49.79 | 23.10 | 41.66 | 42.01 | 20.65 | 40.68 | 15.43 | 7.40 | 15.39 | 35.74 | 17.05 | 32.58 |
| GeoGround (Zhou et al. 2024) | 7B | 58.72 | 25.49 | 46.89 | 53.26 | 29.82 | 48.35 | 13.89 | 4.10 | 15.64 | 41.96 | 19.81 | 36.96 |
| GeoBox-R1 (SFT) | 4B | 55.96 | 30.45 | 45.45 | 51.14 | 27.69 | 45.91 | 22.23 | 15.07 | 19.20 | 43.11 | 24.40 | 36.85 |
| GeoBox-R1 (SFT + GDPO) | 4B | **60.56** | **35.19** | **48.92** | **56.61** | **30.55** | **49.43** | **24.79** | **16.89** | **21.18** | **47.32** | **27.55** | **39.85** |

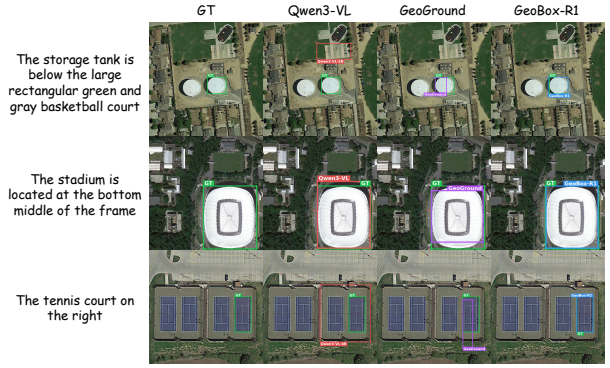



Figure 3: Qualitative HBB grounding comparison among Qwen3-VL, GeoGround, and GeoBox-R1 on representative targets including a storage tank, a stadium, and a tennis court. Green boxes denote ground truth; red, purple, and blue boxes denote the three models' predictions.

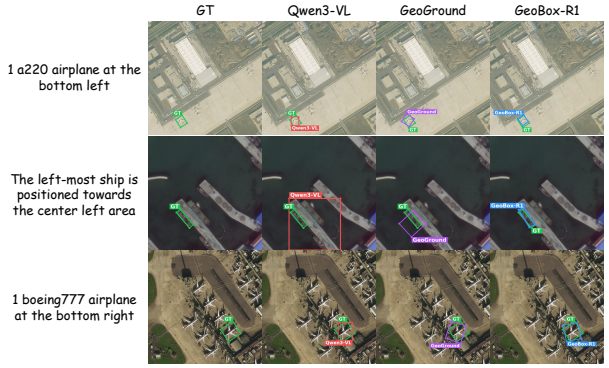



Figure 4: Qualitative OBB comparison on representative airplane and ship targets. When explicitly prompted for OBBs, Qwen3-VL falls back to HBB predictions, whereas GeoGround and GeoBox-R1 produce oriented boxes (colors as in Fig. 3).

and **50.39** mIoU, Compared with GeoGround, the previous published state of the art for unified RSVG, the margins are **+6.43**, **+7.20**, and **+3.59** points, respectively. Among zero-shot baselines, even the strongest general-purpose model Qwen3-VL-8B reaches only 43.70 average Acc@0.5, while InternVL3-2B/8B and LLaVA-OV-1.5-4B/8B obtain 6.85/15.78 and 8.53/11.76. Supervised fine-tuning on refGeo raises the Acc@0.5 averages of LLaVA-OV-1.5-4B and GeoChat to 36.83 and 22.65, and lifts InternVL3-8B to 55.10, making it the strongest baseline in Table 1; yet this backbone, with twice as many parameters and the same SFT recipe, still trails GeoBox-R1 by 3.68 points, suggesting that parameter count alone does not explain the gain. GeoGround leads five of six DIOR-RSVG metrics and all three GeoChat* metrics, whereas GeoBox-R1 leads every metric on RSVG and AVVG. Thus, the macro-average gain reflects complementary per-dataset behavior rather than uniform dominance.

**OBB Results.** As illustrated by Qwen3-VL in Fig. 1, a general-purpose MLLM can fall back to HBB outputs under OBB prompts without task adaptation. GeoBox-R1 achieves the best OBB macro averages on every metric: **47.32** Acc@0.5, **27.55** Acc@0.7, and **39.85** mRIoU, leading the strongest OBB baseline, GeoGround, by 5.36, 7.74, and 2.89 points, respectively. Among the SFT baselines, LLaVA-OV-1.5, GeoChat, and InternVL3-8B reach average Acc@0.5 values of 19.93, 17.00, and 35.74, respectively, leaving a substantial gap to GeoBox-R1. GDPO raises the SFT checkpoint from 43.11 to 47.32 average Acc@0.5. Relative to SFT, the gains are 4.60, 5.47, and 2.56 points on GeoChat*, VRSBench*, and AVVG, respectively. Acc@0.7 and mRIoU also improve on all three benchmarks, indicating tighter geometry across target distributions.

Table 3: Ablation on data organization and CoT under a fixed sample budget (Acc@0.5 macro averages). Without CoT, the 25% CoT split is used as OBB-only supervision. *Mixed* shuffles all samples; *Curriculum-guided* presents them easy-to-hard.

| Data Organization | CoT | HBB @0.5 Avg. | OBB @0.5 Avg. |
|---|---|---|---|
| Curriculum-guided | × | 53.56 | 42.43 |
| Mixed | × | 54.87 | **43.34** |
| Mixed | ✓ | 56.29 | 42.42 |
| Curriculum-guided | ✓ | **57.17** | 43.11 |

## Ablation Studies

**Data Organization and CoT.** Table 3 isolates ordering and CoT under a fixed budget. Without CoT, mixed training leads




Figure 5: GeoBox-R1's HBB-to-OBB CoT reasoning on three challenging car-grounding queries (query, intermediate HBB, refined OBB, GT-vs-prediction overlay, zoomed view). Yellow marks the coarse HBB cue and zoom region; blue is the refined OBB; green is ground truth.

Table 4: Ablation on RL reward design (OBB Acc@0.5, computed with RIoU).

| Model | OBB Acc@0.5 | | | |
|---|---|---|---|---|
| | GeoChat* | VRSBench* | AVVG | Avg. |
| GeoBox-R1 (SFT) | 55.96 | 51.14 | 22.23 | 43.11 |
| + WD | 59.15 | 55.44 | 24.69 | 46.43 |
| + Adaptive WD | 59.04 | 55.26 | 23.80 | 46.03 |
| + RIoU | 59.96 | 56.33 | 24.76 | 47.01 |
| + RIoU + WD | 59.02 | 56.21 | 24.66 | 46.63 |
| + RIoU + Adaptive WD (Default) | **60.56** | **56.61** | **24.79** | **47.32** |

curriculum by 1.31 HBB and 0.91 OBB points. CoT improves HBB by 1.42 points with mixed training and 3.61 with curriculum, reaching 57.17. On OBB, it lowers mixed training from 43.34 to 42.42 but raises curriculum from 42.43 to 43.11. Curriculum-guided CoT therefore leads mixed CoT by 0.88 HBB and 0.69 OBB points, supporting coarse-to-fine training.

**Reward Design.** Table 4 shows that every RL variant improves over SFT by at least 2.92 points, supporting the value of geometric RL. RIoU is the strongest single reward (47.01). Fixed and adaptive WD alone reach 46.43 and 46.03, respectively. The benefit of adaptation appears when paired with the task-aligned overlap signal: RIoU with adaptive WD reaches 47.32, compared with 46.63 for RIoU with fixed WD and 47.01 for RIoU alone. These aggregate results are consistent with adaptive WD serving as a complementary dense signal, although they do not isolate low-overlap or cross-scale subsets.

**Effect of RL beyond OBB.** Although GDPO is trained on OBB samples only, the HBB macro averages rise from 57.17/40.33/48.82 to 58.78/42.22/50.39 for Acc@0.5/Acc@0.7/mIoU. Acc@0.5 improves on six of seven HBB benchmarks, while Acc@0.7 and mIoU improve on all seven; the only Acc@0.5 decrease is 0.26 point on GeoChat*. At the aggregate level, OBB-only RL therefore does not trade away HBB performance and is associated with stronger accuracy at the stricter overlap threshold. This cross-format result is consistent with shared localization knowledge. Small targets remain more sensitive to fixed absolute coordinate errors, especially for OBBs.

## Qualitative Analysis

**HBB Comparison.** In the examples in Fig. 3, Qwen3-VL selects a coarse salient region and GeoGround includes background or drifts toward nearby structures, whereas GeoBox-R1 resolves the storage tank's relation to the reference court, captures the stadium extent tightly, and follows the rightward cue for the tennis court. These cases illustrate differences in target selection, relation understanding, and HBB fitting.

**OBB Comparison.** Fig. 4 shows Qwen3-VL falling back to HBBs and GeoGround producing oriented boxes with corner or angle offsets, while GeoBox-R1 aligns tightly with the elongated ships and airplanes where orientation and aspect ratio are critical.

**CoT Reasoning.** Fig. 5 covers a tiny silver car, dense similar distractors, and the ambiguous query *the widest car*: Step 1 narrows the search with an HBB, Step 2 refines extent and orientation, and the final overlays stay close to ground truth, illustrating the intended coarse-to-fine refinement process.

## Conclusion

We present **GeoBox-R1**, a two-stage paradigm that learns HBB, OBB, and HBB-to-OBB CoT in curriculum order, then refines oriented localization with geometric RL. Across 7 HBB and 3 OBB benchmarks, the 4B model attains the best macro averages on Acc@0.5, Acc@0.7, and mean overlap for both tasks while using fewer parameters than state-of-the-art 7B–8B models, establishing state-of-the-art performance for unified box-level RSVG under refGeo. Ablations support the effectiveness of the CoT curriculum and geometric rewards, and show that OBB-only RL preserves HBB performance while improving oriented localization.

## References


An, X.; Xie, Y.; Yang, K.; Zhang, W.; Zhao, X.; Cheng, Z.; Wang, Y.; Xu, S.; Chen, C.; Zhu, D.; et al. 2025. LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. *arXiv preprint arXiv:2509.23661*.

Baek, W.-K.; Kim, E.; Jeon, H.-K.; Lee, K.-J.; Kim, S.-W.; Lee, Y.-K.; and Ryu, J.-H. 2024. Monitoring Maritime Ship Characteristics Using Satellite Remote Sensing Data from Different Sensors. *Ocean Science Journal*, 59(1): 8.

Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. 2025a. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025b. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum Learning. In *Proceedings of the 26th International Conference on Machine Learning*, 41–48.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. TransVG: End-to-End Visual Grounding with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1769–1779.

El Ghazouali, S.; Gucciardi, A.; Venturini, F.; Venturi, N.; Rueegsegger, M.; and Michelucci, U. 2024. FlightScope: An Experimental Comparative Review of Aircraft Detection Algorithms in Satellite Imagery. *Remote Sensing*, 16(24): 4715.

Gong, J.; Liu, C.; and Huang, X. 2020. Advances in Urban Information Extraction from High-Resolution Remote Sensing Imagery. *Science China Earth Sciences*, 63(4): 463–475.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645: 633–638.

Hou, L.; Lu, K.; Yang, X.; Li, Y.; and Xue, J. 2023. G-Rep: Gaussian Representation for Arbitrary-Oriented Object Detection. *Remote Sensing*, 15(3): 757.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, S.; He, S.; and Wen, B. 2026. RSGround-R1: Rethinking Remote Sensing Visual Grounding through Spatial Reasoning. *arXiv preprint arXiv:2601.21634*.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27831–27840.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. LISA: Reasoning Segmentation via Large Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Li, K.; Wang, D.; Xu, H.; Zhong, H.; and Wang, C. 2024. Language-Guided Progressive Attention for Visual Grounding in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–13.

Li, X.; Ding, J.; and Elhoseiny, M. 2024. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. *Advances in Neural Information Processing Systems*, 37: 3229–3242.

Li, Z.; Muhtar, D.; Gu, F.; He, Y.; Zhang, X.; Xiao, P.; He, G.; and Zhu, X. 2025. LHRS-Bot-Nova: Improved Multimodal Large Language Model for Remote Sensing Vision-Language Interpretation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227: 539–550.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, J.; Chen, Q.; Wang, Z.; Tang, Y.; Zhang, Y.; Yan, C.; Wang, D.; Li, X.; and Zhao, B. 2025a. AerialVG: A Challenging Benchmark for Aerial Visual Grounding by Exploring Positional Relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5177–5187.

Liu, S.-Y.; Dong, X.; Lu, X.; Diao, S.; Belcak, P.; Liu, M.; Chen, M.-H.; Yin, H.; Wang, Y.-C. F.; Cheng, K.-T.; et al. 2026. GDPO: Group reward-Decoupled Normalization Policy Optimization for Multi-reward RL Optimization. *arXiv preprint arXiv:2601.05242*.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025b. Visual-RFT: Visual Reinforcement Fine-Tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2034–2044.

Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; and Xiao, P. 2024. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. In *European Conference on Computer Vision*, 440–457.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824*.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. GLaMM: Pixel Grounding Large Multimodal Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13009–13018.

Shabbir, A.; Zumri, M.; Bennamoun, M.; Khan, F. S.; and Khan, S. 2025. GeoPixel: Pixel Grounding Large Multimodal Model in Remote Sensing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 54095–54111.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*.

Sun, Y.; Feng, S.; Li, X.; Ye, Y.; Kang, J.; and Huang, X. 2022. Visual Grounding in Remote Sensing Images. In *Proceedings of the 30th ACM International Conference on Multimedia*, 404–412.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, S.; Jin, S.; Zhang, W.; Xu, L.; Liu, W.; Li, W.; and Loy, C. C. 2025. F-LMM: Grounding Frozen Large Multimodal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24710–24721.

Yang, X.; Yang, X.; Yang, J.; Ming, Q.; Wang, W.; Tian, Q.; and Yan, J. 2021. Learning High-Precision Bounding Box for Rotated Object Detection via Kullback-Leibler Divergence. *Advances in Neural Information Processing Systems*, 34: 18381–18394.

Yang, X.; Zhang, G.; Yang, X.; Zhou, Y.; Wang, W.; Tang, J.; He, T.; and Yan, J. 2023. Detecting Rotated Objects as Gaussian Distributions and its 3-D Generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4335–4354.

Yang, Y.; He, X.; Pan, H.; Jiang, X.; Deng, Y.; Yang, X.; Lu, H.; Yin, D.; Rao, F.; Zhu, M.; et al. 2025. R1-Onevision: Advancing Generalized Multimodal Reasoning through Cross-Modal Formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2376–2385.

Ye, J.; Tian, J.; Yan, M.; Yang, X.; Wang, X.; Zhang, J.; He, L.; and Lin, X. 2022. Shifting More Attention to Visual Backbone: Query-Modulated Refinement Networks for End-to-End Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15502–15512.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2023. RSVG: Exploring Data and Models for Visual Grounding on Remote Sensing Data. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Zhan, Y.; Xiong, Z.; and Yuan, Y. 2025. SkyEyeGPT: Unifying Remote Sensing Vision-Language Tasks via Instruction Tuning with Large Language Model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221: 64–77.

Zhang, J.; Huang, J.; Yao, H.; Liu, S.; Zhang, X.; Lu, S.; and Tao, D. 2025. R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1859–1869.

Zhang, P.; Zhang, Y.; Xu, L.; Lin, J.; Guo, Z.; Wang, F.; Yang, X.; Wei, K.; and Wang, L. 2026. GeoViS: Geospatially Rewarded Visual Search for Remote Sensing Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14335–14345.

Zhang, Y.; Ma, Z.; Gao, X.; Shakiah, S.; Gao, Q.; and Chai, J. 2024. GROUNDHOG: Grounding Large Language Models to Holistic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14227–14238.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; et al. 2025. SWIFT: A Scalable Lightweight Infrastructure for Fine-Tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 29733–29735.

Zhou, Y.; Lan, M.; Li, X.; Feng, L.; Ke, Y.; Jiang, X.; Li, Q.; Yang, X.; and Zhang, W. 2024. GeoGround: A Unified Large Vision-Language Model for Remote Sensing Visual Grounding. *arXiv preprint arXiv:2411.11904*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*.

Zong, Y.; Zhang, Q.; An, D.; Li, Z.; Xu, X.; Xu, L.; Tu, Z.; Xing, Y.; and Dabeer, O. 2025. Ground-V: Teaching VLMs to Ground Complex Instructions in Pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24635–24645.