Beyond Referring Expressions: Scenario Comprehension Visual Grounding

Ruozhen He[1], Nisarg A. Shah[2], Qihua Dong[3], Zilin Xiao[1], Jaywon Koo[1], Vicente Ordonez[1]
[1]Rice University [2]Johns Hopkins University [3]Northeastern University
{catherine.he,zilin,jk125,vicenteor}@rice.edu
nshah82@jhu.edu dong.qih@northeastern.edu

Abstract

Existing visual grounding benchmarks primarily evaluate alignment between image regions and literal referring expressions, where models can often succeed by matching a prominent named category. We explore a complementary and more challenging setting of scenario-based visual grounding, where the target must be inferred from roles, intentions, and relational context rather than explicit naming. We introduce Referring Scenario Comprehension (RSC), a benchmark designed for this setting. The queries in this benchmark are paragraph-length texts that describe object roles, user goals, and contextual cues that include deliberate references to distractor objects that often require deep understanding to resolve. Each instance is annotated with interpretable difficulty tags for uniqueness, clutter, size, overlap, and position which expose distinct failure modes and support fine-grained analysis. RSC contains approximately 31k training examples, 4k in-domain test examples, and a 3k out-of-distribution split with unseen object categories. We further propose ScenGround, a curriculum reasoning method serving as a reference point for this setting, combining supervised warm-starting with difficulty-aware reinforcement learning. Experiments show that scenario-based queries expose systematic failures in current models that standard benchmarks do not reveal, and that curriculum training improves performance on challenging slices and transfers to standard benchmarks.

1 Introduction

Visual grounding, the task of associating an image region with a natural language description, is a core capability for embodied assistants and multimodal models (Geng et al., 2025; Shen et al., 2021; Rajabi and Kosecka, 2024). Despite rapid progress in building capable vision-and-language models (VLMs), strong performance on standard referring comprehension (REC) often relies on explicit lexical cues, particularly category references and salient attributes that can be directly matched to candidate regions without requiring deep language understanding (Yu et al., 2016; Mao et al., 2016; Liu et al., 2017; Arbelle et al., 2021; Yang et al., 2023). This protocol creates a systematic blind spot as models optimized for literal phrase matching may fail when users express a need for an object by describing situations instead of making explicit object references.

Real referring behavior is not necessarily concise and direct. Natural language allows users to describe a target through a situational need, an object role, or a user goal. For example, a user attempting to find what time it is may prompt a model on *"checking the time"* instead of explicitly asking for *"the clock"*. Such scenario-based queries might contain multiple disambiguating cues, demanding reasoning over intent and visual context, not just category lookup. Despite being a natural and valid form of reference, this type of queries is absent from existing grounding benchmarks, leaving a gap in how we evaluate and develop grounding models. As illustrated in Figure 1, the same target object are described in completely different ways under traditional REC and scenario-based grounding: the former can succeed via lexical matching, while the latter requires integrating relational, spatial, and intentional cues while actively ignoring explicitly mentioned distractors.

We introduce **Referring Scenario Comprehension (RSC)**, a benchmark designed to study this under-examined setting. RSC replaces referring phrases with scenario-based queries that describe a user role, goal, and at least three disambiguating cues, and deliberately mentions competing objects to require deep understanding. Each instance is annotated with reasoning traces and five interpretable difficulty tags (**U**niqueness, **C**lutter, **S**ize, **O**verlap, and **P**osition), which expose distinct fail-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 40742–40762
October 24-29, 2026 ©2026 Association for Computational Linguistics

	Traditional REC	Referring Scenario Comprehension (RSC)	RSC Tags
	Q: the clock behind the pole **A:** [292,381,115,112]	**Q:** As a tourist trying to catch a bus, you want to check the time. You look for the round, silver-framed object with black markings, mounted on a pole, partly hidden by a railing, and positioned above the sidewalk near the street, not to be confused with car wheels or headlights. **A:** {"target_object": clock, "bbox": [292,381,115,112]}	**U**niqueness: Yes **C**lutter: Medium **S**ize: Large **O**verlap: Low **P**osition: Center
	Q: the yellow-orange apple with a red patch **A:** [426,316,194,164]	**Q:** As a wildlife photographer documenting animal feeding habits, you want a close-up of the fruit that is close to the feet of the bird. You need the partially eaten, yellow-red object, positioned lower right, directly in front of the bird's beak, and distinct from any whole fruit or background elements. **A:** {"target_object": apple, "bbox": [426,316,194,164]}	**U**niqueness: Yes **C**lutter: Low **S**ize: Large **O**verlap: Low **P**osition: Off-Center
	Q: the brown leather glove held by the catcher **A:** [145,122,75,63]	**Q:** As a coach reviewing game footage, you want to analyze how the crouching player's equipment helps catch fast pitches. Focus on the large brown item on their left hand, which is directly in the ball's path, below the batter's swinging bat, and distinct from the black chest protector and helmet. **A:** {"target_object": baseball glove, "bbox": [145,122,75,63]}	**U**niqueness: No **C**lutter: High **S**ize: Medium **O**verlap: High **P**osition: Center

Figure 1: **Referring Scenario Comprehension (RSC) vs. Referring Expression Comprehension (REC).** Each row shows a target object under both settings. REC queries often name the target category directly, allowing success via lexical matching. RSC instead pairs each image with a lengthy scenario-based query specifying a user role, goal, and multiple disambiguating cues, including explicit contrasts against competing objects, and requires output identifying both the target object and its bounding box. The RSC difficulty tags (U/C/S/O/P: Uniqueness, Clutter, Size, Overlap, Position) characterize each instance, enabling fine-grained training and evaluation.

ure modes and support fine-grained curriculum design and evaluation. RSC contains 31k training examples, 4k in-domain test examples, and a 3k out-of-distribution split with unseen object categories, enabling evaluation of both in-domain disambiguation and cross-category generalization.

Table 1 situates RSC among existing visual grounding benchmarks. Most prior resources contain short literal phrases, are annotated via crowd-sourced expression games or scene-graph templates, and provide no out-of-distribution (OOD) evaluation split. Two recent works move toward richer queries: SK-VG (Chen et al., 2023b) requires reasoning over external knowledge paragraphs, and EgoIntention (Sun et al., 2025) targets implicit affordance grounding in egocentric images. RSC addresses a complementary gap, exocentric scenario grounding, where the challenge is reasoning over rich paragraph-length scenarios and disambiguating among visually similar instances using rich relational and intentional cues rather than inferring affordances or external knowledge. RSC also provides an OOD split, multi-axis difficulty tags, and per-instance reasoning trace annotations to support diverse learning and analysis.

Beyond introducing the benchmark, we also propose a strong baseline **ScenGround**, a two-stage curriculum reasoning method for scenario-based visual grounding. In Stage 1, *Thought-Primed SFT (TP-SFT)* aligns the model to the output schema and elicits faithful reasoning traces before a structured answer, using the easier RSC slices to stabilize interface learning. In Stage 2, *Incentive-Curriculum GRPO (IC-GRPO)* refines localization and disambiguation via shaped rewards coupling geometry, including smooth IoU reward with center-consistency and out-of-bounds penalties, and alias-aware category rewards. The training follows a tag-aware curriculum, feeding more difficult non-unique, cluttered, overlapping, and off-center targets in the later stage. A prompt-template ensemble further improves robustness across query surface forms. ScenGround presents a well-characterized baseline demonstrating that difficulty-aware curriculum training substantially improves scenario grounding and transfers the improvement to standard benchmarks.

Our main contributions are as follows:

- We introduce **RSC**, a scenario-based visual grounding benchmark with difficulty-controlled instances, per-instance reasoning trace annotations, and an OOD split with disjoint object categories.
- We propose **ScenGround**, a curriculum

Table 1: **Comparison of referring expression grounding benchmarks.** We compare existing datasets with RSC across dataset scale, evaluation splits, referring/query style, reasoning supervision, difficulty annotations, and competing-object mentions. Most prior datasets rely on short literal phrases and lack explicit reasoning traces or OOD evaluation. RSC introduces scenario-based queries written as natural language paragraphs and provides per-instance reasoning trace annotations. It further labels five interpretable difficulty factors (U/C/S/O/P) and includes both in-distribution (ID) and out-of-distribution (OOD) category test sets to evaluate reasoning and generalization.

Benchmark	Train	ID Test	OOD Test	Referring Style	Query Style	Avg. $\lvert q \rvert$	Reasoning Traces	Difficulty Tags	Competing Mentions
RefCOCO (Kazemzadeh et al., 2014)	120,624	10,752	✗	Literal	Phrase	3.5	✗	✗	✗
RefCOCO+ (Kazemzadeh et al., 2014)	120,191	10,615	✗	Literal	Phrase	3.5	✗	✗	✗
RefCOCOg (Mao et al., 2016)	80,512	9,602	✗	Descriptive literal	Phrase	8.2	✗	✗	✓
Cops-Ref (Chen et al., 2020)	119,603	12,586	✗	Compositional literal	Template	14.4	✗	✗	✓
Ref-Reasoning (Yang et al., 2020)	721,164	34,609	✗	Literal	Template	8.5	Graph	Graph layout	✓
SK-VG (Chen et al., 2023b)	23,403	6,597	✗	Scene knowledge Q&A	Q&A	5.3	✗	E/M/H	✓
FineCops-Ref (Liu et al., 2024a)	163,792	9,605	✗	Compositional literal	Phrase	12.2	✗	1/2/3	✓
EgoIntention (Sun et al., 2025)	15,667	9,892	✗	Egocentric intention query	Sentence	18.8	✗	✗	✗
RSC (Ours)	**31,342**	**4,038**	**3,247**	**Scenario-based query**	Paragraph	**52.7**	Paragraph	U/C/S/O/P	✓

reasoning method combining supervised warm-starting with difficulty-aware reinforcement learning, providing a strong and well-characterized reference point for scenario-based grounding.

- Experiments demonstrate that RSC exposes failure modes invisible to standard benchmarks, and that difficulty-aware curriculum training transfers improvements to standard referring benchmarks.

2 Related Work

Referring Expression Datasets and Methods. Referring expression aims to localize image regions described by natural language. Early datasets (Yu et al., 2016; Kazemzadeh et al., 2014; Mao et al., 2016; Plummer et al., 2015) introduced the task using short, literal phrases that directly name the target objects. Subsequent benchmarks expanded the scope along multiple axes: compositional reasoning with hard negatives (Liu et al., 2019; Chen et al., 2020; Liu et al., 2024a; Dong et al., 2026), structured reasoning over scene graphs (Yang et al., 2020), segmentation-based grounding (Lai et al., 2024; Wu et al., 2020), 3D referring expressions (Chen and Chang, 2020; Achlioptas et al., 2020) and GUI domains (You et al., 2024). More recently, SK-VG (Chen et al., 2023b) introduces grounding conditioned on long-form scene knowledge, and EgoIntention (Sun et al., 2025) targets egocentric intention grounding where models must infer the user intent from first-person views. To the best of our knowledge, no existing benchmark evaluates grounding from scenario-based queries: paragraph-length descriptions specifying user roles, goals, and explicit distractor contrasts, where the challenge lies in reasoning over rich intentional context to disambiguate distractors.

Visual Grounding with VLMs. Early visual grounding methods adopted modular pipelines pairing region proposals with language encoders (Kamath et al., 2021; Li et al., 2022). Later transformer-based methods reformulated referring expression as an end-to-end task (Deng et al., 2021; Yang et al., 2023; He et al., 2024; Zhan et al., 2025). Grounding DINO (Liu et al., 2024b) further unified open-set object detection with text-conditioned localization. With the rise of large vision-language models (LVLMs), grounding has become an inherent capability within general-purpose multimodal systems. Models such as Ferret (You et al., 2023), Shikra (Chen et al., 2023a), Kosmos-2 (Peng et al., 2023), PEVL (Yao et al., 2022), and SCoRD (Yang et al., 2024) used region-level referring and grounding in LVLMs by representing spatial coordinates as part of the text generation process. Subsequent work scaled these ideas: Qwen-VL (Wang et al., 2024) and Qwen2.5-VL (Bai et al., 2025a) align image–caption–box tuples during pretraining to support open-vocabulary grounding, while InternVL (Chen et al., 2024c,b) and VisionLLM (Wu et al., 2024) extend grounding to hundreds of vision-language tasks. Most recently, UniVG-R1 (Bai et al., 2025b) incorporates reinforcement learning with Chain-of-Thought reasoning (Wei et al., 2022) for universal grounding. Despite these advances, existing methods are mainly evaluated

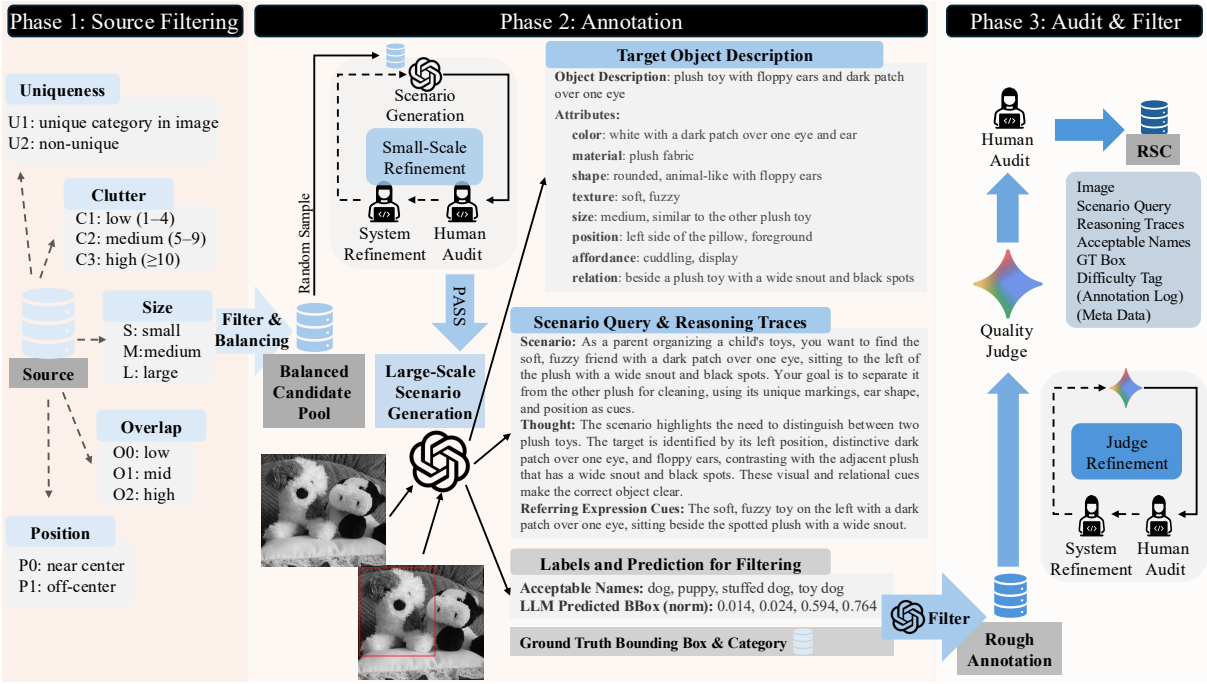


Figure 2: **Phase 1** filters and balances source instances computing five interpretable difficulty tags to form a tag-balanced candidate pool. **Phase 2** generates annotations via a two-stage process: a small-scale refinement loop first validates the generation prompt through iterative system refinement and human audit, before large-scale scenario generation is applied to the full candidate pool. Each instance is annotated with a target object description, a scenario query with reasoning traces, acceptable category aliases, and an LLM-predicted bounding box for quality filtering. **Phase 3** applies automatic and human quality control: a quality judge filters rough annotations, human auditors verify a stratified sample. The final RSC dataset provides, per instance, a scenario query, reasoning traces, acceptable names, ground-truth box, and difficulty tags.

on short object-centric queries, and the scenario-based grounding setting, where success requires reasoning over roles, goals, and relational context, remains overlooked. RSC is designed to fill this gap, and ScenGround provides strong curriculum reasoning tailored for this setting.

3 Referring Scenario Comprehension

Referring Scenario Comprehension (RSC) asks a model to localize an object from a *scenario* which describes a user role, a goal, distractors, and non-literal disambiguating cues. Given an image $x \in \mathbb{R}^{H\times W\times 3}$ and a scenario s, a model ϕ predicts:

$$\phi(x, s) \;\rightarrow\; (\widehat{y}, \widehat{b}), \quad \widehat{b} = (\hat{x}, \hat{y}, \hat{w}, \hat{h}) \in \mathbb{Z}^4_{\geq 0}, \tag{1}$$

where $\widehat{b}$ is an xywh box clipped to $[0, W] \times [0, H]$. The dataset provides tuples

$$\mathcal{D} = \{(x_i,\, s_i,\, y_i,\, b_i,\, \mathcal{A}_i,\, e_i,\, r_i,\, \tau_i)\}_{i=1}^{N}, \tag{2}$$

with gold category y_i, box b_i, acceptable aliases $\mathcal{A}_i$, a concise referring expression e_i, a reasoning trace r_i, and interpretable difficulty tags $\tau_i = (U_i, C_i, S_i, O_i, P_i)$ for **U**niqueness, **C**lutter, **S**ize, **O**verlap, and **P**osition. By construction, s_i does not reveal y_i. The curation pipeline proceeds in three phases, illustrated in Figure 2.

Phase 1: Source Filtering and Balancing. We source instances from MS-COCO (Lin et al., 2014), preferring images that overlap with RefCOCO/+/g for visual recognizability, and prevent leakage by excluding evaluation image IDs, retaining category labels only as hidden construction signals, and de-duplicating by image hash and annotation ID before splitting. The OOD split draws from LVIS (Gupta et al., 2019) with COCO overlaps and synonym collisions removed to ensure strict category disjointness at both string and synset levels.

For each instance i with box $b_i = (x_i, y_i, w_i, h_i)$ and center c_i, we compute:

$$a_i = \frac{w_i h_i}{HW},\; d_i = \frac{\|c_i - (\frac{W}{2}, \frac{H}{2})\|_2^2}{W^2 + H^2},\; o_i = \sum_{j\neq i} \mathrm{IoU}(b_i, b_j). \tag{3}$$

Difficulty tags are assigned by quantile binning estimated on the candidate pool *before* splitting: *Size (S)* via tertiles of a_i; *Overlap (O)* via low/mid/high percentiles of o_i; *Position (P)* via a median split of d_i; *Uniqueness (U)* by the pres-

ence of same-category distractors (m_i=0 ⇒U1; $m_i \geq 1$ ⇒U2); and *Clutter (C)* by binning the per-image instance count N_{img}. These axes expose ambiguity (U), scene density (C), scale (S), occlusion and congestion (O), and off-center placement (P) in a way that is both interpretable and controllable.

To form a balanced candidate pool, we target split-wise marginal proportions $\Pi = \{\pi_U, \pi_C, \pi_S, \pi_O, \pi_P\}$ and enforce per-category quotas to avoid category dominance. Allocation proceeds hierarchically: (i) assign a quota per category using long-tail-favoring priorities; (ii) partition each category's quota across tag combinations according to Π; (iii) sample within each bin. A continuous difficulty score $D_i \in [0, 1]$ uses a monotone combination of quantile-normalized (a_i, o_i, d_i), distractor count, and category rarity, ordering instances within each bin.

Phase 2: Annotation. Inspired by prior model-in-the-loop pipelines (Kirillov et al., 2023; Wang et al., 2023; Xu et al., 2025), our scenario generation follows a two-stage process to ensure annotation quality before scaling. In a *small-scale refinement loop*, we first generate scenarios for a small set of random samples, then conduct iterative system refinement and a human audit to validate and improve the generation prompt. Generation proceeds to large scale only once the audited pass rate exceeds a quality threshold of 90%.

In *large-scale generation*, an LLM receives two views of each target image: the full image and the same image with a red rectangle prior marking the target region (the rectangle is not part of the object). It then returns a structured JSON containing: a concise category-free referring expression ($\leq$25 words), a user-driven scenario (role, goal, and $\geq$3 disambiguating cues from attributes, relations, position, and affordance), a reasoning trace explaining how the scenario maps to visual evidence, canonical object attributes, an alias set $\mathcal{A}_i$, and a predicted bounding box $\hat{b}_i$. The prompt enforces that category names are absent from the scenario and expression, and requires at least one explicit contrast with a plausible distractor.

Phase 3: Audit and Filter. Instances passing the automatic gates in Phase 2 form a rough annotation pool, which undergoes dual-track quality control summarized in Table 2. The LLM GT filter, comprising schema validation, box IoU gate, alias consistency, and leakage detection, retains 98–99% of ID instances and 79.7% of OOD instances. The lower OOD retention reflects the greater visual ambiguity of LVIS categories, for which the LLM more frequently fails to produce an accurate box or alias set. This confirms the filter is performing meaningful quality control on harder categories rather than passing annotations indiscriminately.

Table 2: **RSC data retention across quality control stages.** After *LLM GT Filter*: instances passing the automatic schema, box IoU, alias consistency, and leakage gates. After *Quality Filter*: instances passing the full judge-and-audit pipeline.

Split	Initial	GT Filter	Kept	Quality Filter	Kept
Train (SFT)	30,000	29,551	98.5%	23,802	80.5%
Train (RL)	10,000	9,884	98.8%	7,540	76.3%
Test (ID)	5,000	4,915	98.3%	4,038	82.2%
Test (OOD)	5,000	3,983	79.7%	3,247	81.5%

Instances passing the LLM GT filter then enter the quality filter stage. A *quality judge* automatically scores each annotation for the uniqueness of the scenario referring target, and the accuracy of the bounding box for the scenario. Borderline cases enter a *judge refinement* loop applying targeted system and human correction before re-scoring. After this stage, 76–82% of LLM-filtered instances are retained across splits, yielding final counts of 23,802 training (SFT), 7,540 training (RL), 4,038 in-domain test, and 3,247 OOD test examples.

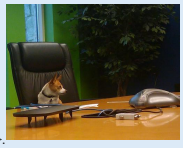
In parallel, three independently drawn samples of 100 instances each undergo *human audit* by three expert annotators, who independently verify scenario non-leakage, alias consistency ($y_i \in \mathcal{A}_i$), box correctness ($\text{IoU}(\hat{b}_i, b_i) \geq \theta_{\text{bbox}}$), and attribute faithfulness. Majority-vote accuracy was 95.7% overall (per-rater: 94%/96%/97%), with substantial agreement (Fleiss' κ= *0.94*). The residual 4% errors were attributable to LLM-hallucinated distractor descriptions ($\sim$3%) and minor attribute misdescriptions ($\sim$1%).

Each released RSC instance provides: a scenario query, reasoning traces, acceptable category names, ground-truth bounding box, and difficulty tag, with annotation logs and metadata included for research use. *More details are provided in the appendix.*

4 ScenGround

ScenGround is a two-stage curriculum reasoning method for scenario-based grounding on RSC. Thought-Primed SFT (TP-SFT) aligns the interface and elicits faithful `<think>` traces that precede `<answer>`. Incentive-Curriculum GRPO (IC-GRPO) then refines reasoning and localization by optimizing shaped rewards over RSC's difficulty-stratified curriculum.

You are given an image and a referring scenario description.
First, think step by step inside <think>...</think>.
Then output ONE JSON object inside <answer>...</answer> exactly:
<answer>{
"target_object": "<canonical object category>",
"bbox": [x, y, w, h]
}</answer>
Rules:
- The only content outside <answer> should be <think>…</think>.
- "bbox" are 4 INTEGERS [x,y,w,h], xywh in pixels, clipped to bounds.
Image size: Height=427, Width=640
Referring scenario description:
As a meeting organizer checking the room, you want to photograph the unexpected guest. You need the small, brown-and-white animal sitting on the leftmost chair, with pointed ears and a collar, ignoring the conference phone and cables on the table.



<think>The scenario singles out the only animal present, described by its brown-and-white fur, small size, and upright posture on the left chair. The contrast with the inanimate conference phone and cables, plus the unique position and pointed ears, makes identification unambiguous. The target object is that small brown-and-white animal sitting on chair.</think>
<answer>{"target_object": "dog", "bbox": [88, 200, 86, 100]}</answer>

Figure 3: **ScenGround prompt and output schema.** Given an image and a user-driven scenario, the model is instructed to reason inside `<think>` and to emit a structured JSON inside `<answer>`. The JSON contains `target_object` and `bbox` `[x,y,w,h]`. Scenarios avoid category names and force disambiguation via attributes, relations, and spatial cues, while `<think>` must justify the selection and ignore distractors. This schema is used in both TP-SFT and IC-GRPO. To increase the robustness of ScenGround, IC-GRPO uses 7 other templates.

4.1 Thought-Primed SFT

Given image x, scenario s, and target text $\mathbf{y}$ (the concatenation of `<think>` and `<answer>` spans), we optimize the standard next-token loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,s,\mathbf{y})\sim\mathcal{D}_{\text{sft}}}\left[\sum_{t=1}^{|\mathbf{y}|}\log p_\theta(y_t \mid x, s, \mathbf{y}_{<t})\right]. \tag{4}$$

The output schema requires a single JSON inside `<answer>` with keys `target_object` and `bbox` (xywh, pixel integers). Training uses the easier RSC slices (D_i easy percentiles) to stabilize schema learning before RL. TP-SFT teaches the output schema, elicits faithful reasoning traces inside `<think>`, and provides a stable reference policy π_{ref} for the subsequent RL stage.

4.2 Incentive-Curriculum GRPO

We fine-tune π_θ using GRPO (Guo et al., 2025) with group-relative advantages (Eq. 6) and a KL-regularized objective with adaptive β tracking a target KL band throughout training (see Appendix D).

Shaped rewards. The scalar reward combines four components (Eq. 13). The *geometry reward* r_{iou} combines a base IoU term with smooth logistic bonuses near two operating points, a small center-consistency term, and a penalty for out-of-bounds predictions. The *category reward* r_{cat} is alias-aware: it awards full credit for canonical names, partial credit for accepted aliases and token-overlap matches, and is gated by a minimum IoU threshold to discourage well-labelled but poorly localized predictions. The *format* and *structure rewards* enforce schema compliance, rewarding parseable JSON inside `<answer>` with the required keys and penalizing malformed outputs. Reward weights are linearly annealed to increase geometry emphasis in later training.

Tag-aware curriculum. IC-GRPO samples from RSC using difficulty scores D_i. Stage 1 draws predominantly from easy-to-medium slices; Stage 2 shifts toward harder instances with non-unique categories (U2), high clutter (C3), high overlap (O2), and off-center placement (P1). This progressive schedule addresses reward sparsity: easy instances first establish reliable IoU signals before the policy encounters harder disambiguations where the category reward requires a cleared IoU gate.

Prompt-template ensemble. To reduce sensitivity to surface-level query phrasing, we uniformly sample from eight prompt paraphrases (PTE-8) per training step. All templates share the same output schema; rewards are logged per template for analysis but supervision is identical, improving robustness without changing the learning objective.

5 Experimental Setup

Implementation details. Annotation is generated by GPT4o (Hurst et al., 2024) and the quality judge relies on Gemini-2.5-Pro (Comanici et al., 2025). We train ScenGround on Qwen2.5-VL-7B-Instruct (Bai et al., 2025a) in two stages. TP-SFT uses AdamW with learning rate 5×10^{-6} for 5 epochs. GRPO Stage 1 uses AdamW (lr 1×10^{-6}) for 5 epochs with K=6 rollouts per prompt; Stage 2 uses lr 2×10^{-6} for 2 epochs with K=12 rollouts, generation batch 288, temperature 0.9, top-p 0.92, and max completion length 160. The KL coefficient β is initialized at 2×10^{-2} and adapted online in both stages. All remaining hyperparameters are provided in Appendix H.

Data for curriculum learning. We partition the raw instances at the image level into disjoint sets $\mathcal{S}_{\text{SFT}}$, $\mathcal{S}_{\text{RL}}$, and $\mathcal{S}_{\text{test}}$, enforcing no image leakage across splits. After RSC curation filters, we retain 23k instances for SFT and 7k for RL. Difficulty is binned by global quantiles of D_i into easy, medium, and hard buckets. The SFT split targets an easy-purity of at least 0.70. RL Stage 1 uses mixture $\Pi^{(1)}$=(0.70, 0.30, 0.00) and Stage 2 shifts to $\Pi^{(2)}$=(0.20, 0.60, 0.20) for (easy, medium, hard).

Table 3: **Performance on RSC** (ID and OOD). Metrics: mIoU and $\mathrm{Acc}@\{0.5, 0.7\}$ (higher is better); Cat Acc = category accuracy. ‡**Oracle settings, not directly comparable:** Grounding DINO receives privileged inputs unavailable at inference: *cat token* provides the gold category name directly and *ref. cue* feeds the conciser reasoning trace conclusion "referring expression cue" (See Figure 2) as the open-vocabulary query.

	RSC-ID				RSC-OOD			
Model	mIoU	Acc@0.5	Acc@0.7	Cat Acc	mIoU	Acc@0.5	Acc@0.7	Cat Acc
GPT-4o (Hurst et al., 2024)	19.41	13.23	5.37	79.45	16.57	9.55	3.08	**62.00**
Claude 3.7 (Anthropic, 2025)	16.64	8.32	3.71	89.67	12.04	5.54	1.87	58.98
Grounding DINO (*cat token*)‡	44.60	47.55	42.03	—	32.18	31.99	27.89	—
Grounding DINO (*ref. cue*)‡	48.99	51.84	46.02	—	38.12	38.26	34.07	—
InternVL2.5 8B (Chen et al., 2024a)	16.76	11.88	6.74	81.70	8.08	3.64	1.61	36.50
Qwen3-VL 8B (Team, 2025)	15.46	11.17	6.05	75.04	7.38	3.70	1.48	46.97
Qwen2.5-VL 7B (Bai et al., 2025a)	30.31	27.42	15.66	30.86	21.54	15.88	9.19	20.82
ScenGround (Ours)	55.68	60.90	42.32	94.23	38.37	38.11	22.64	21.13

Table 4: **Ablation of curriculum in IC-GRPO** on RSC in-domain (RSC-ID) and out-of-domain (RSC-OOD). Metrics: mIoU and $\mathrm{Acc}@\{0.5, 0.7\}$ for boxes; Cat Acc for category naming. *Single Stage* trains on the union of all RL samples without curriculum; *Stage-k* trains on the stage-k slice, all other settings unchanged.

	RSC-ID				RSC-OOD			
Method	mIoU	Acc@0.5	Acc@0.7	Cat Acc	mIoU	Acc@0.5	Acc@0.7	Cat Acc
Qwen2.5-VL 7B (Bai et al., 2025a)	30.31	27.42	15.66	30.86	21.54	15.88	9.19	20.82
SFT	55.01	60.51	42.59	89.05	33.20	34.03	20.36	12.53
GRPO, Single Stage	54.04	59.04	38.43	94.90	37.66	36.99	20.82	20.05
GRPO, Stage 1	**55.68**	60.95	41.88	94.04	36.93	36.59	21.37	19.00
GRPO, Stage 2	**55.68**	60.90	**42.32**	**94.23**	**38.37**	**38.11**	**22.64**	**21.13**

Table 5: **Referring expression comprehension** on RefCOCO+ and RefCOCOg. Metric: Acc@0.5 (%, ↑). † Use the official grounding pipeline and prompt (Bai et al., 2025a) . *Use a custom prompt.

	RefCOCO+			RefCOCOg	
Method	Val	Test A	Test B	Val	Test
Qwen2.5-VL	84.20	89.10	76.90	87.20	87.20
Qwen2.5-VL*	52.54	56.75	52.53	52.46	51.40
ScenGround*, SFT	50.86	43.84	50.74	65.45	63.39
ScenGround*, GRPO	70.16	74.63	70.05	78.19	75.61

6 Experiments

Comparison on Referring Scenario Comprehension. Table 3 evaluates two closed-source LLMs (Hurst et al., 2024; Anthropic, 2025), one open-vocabulary grounding model (Liu et al., 2024b), three open-source LVLMs (Bai et al., 2025a; Team, 2025; Chen et al., 2024a) on RSC. The results reveal a consistent pattern across baselines: models with strong category accuracy tend to lag on localization, while strong detectors lack semantic reasoning. This trade-off reflects the core challenge RSC is designed to expose: scenario grounding requires jointly strong spatial localization and scenario-aware semantic inference, a combination underemphasized in prior benchmarks.

The Grounding DINO results illustrate the gap between detection-style grounding and full scenario reasoning. Under the oracle *cat token* setting, which directly supplies the gold category name, Grounding DINO achieves strong localization. Under the *ref. cue* setting, which supplies a short descriptive non-literal phrase rather than the full scenario, performance improves further, particularly on OOD. These gains confirm that Grounding DINO's architecture is well-optimized for direct-cue detection, but neither setting reflects the full RSC protocol given to other models.

Among models that receive only the full scenario query s and must jointly predict the target category and bounding box, ScenGround substantially outperforms all baselines on ID mIoU. Closed-source models achieve high category accuracy but poor localization, confirming that accurate understanding does not imply performant spatial grounding. The

localization and target category prediction gains from ScenGround are consistent across ID and OOD splits, reducing the baseline trade-off and delivering robust performance on RSC.

Effectiveness of GRPO and Curriculum Learning. Table 4 ablates training stages. TP-SFT already delivers a large ID jump over the off-the-shelf VLM (Bai et al., 2025a), showing that aligning the model with the reasoning schema and the RSC objective teaches strong box prediction. However, OOD category accuracy remains the dominant failure mode after SFT. Moving to IC-GRPO without curriculum (*Single Stage*) improves OOD across both localization and category naming, but slightly regresses ID localization relative to SFT. This is consistent with reward sparsity: when easy and hard cases are mixed, many early trajectories fall below the IoU gate that modulates the alias reward, yielding high-variance gradients and modest gains.

The tag-aware curriculum addresses this. *Stage 1* (easy to medium) preserves the strong ID accuracy of SFT and improves OOD over SFT, but falls short of Stage 2. *Stage 2* (medium to hard) yields the best results across all OOD metrics while keeping ID performance essentially unchanged. As geometric accuracy improves on harder, more ambiguous instances, the policy learns to map scenario cues to both *where* and *what* under category shift. However, ScenGround's OOD category accuracy remains only marginally above the untuned baseline, indicating that cross-category semantic naming is largely unsolved by curriculum training alone. This may be attributed to the fine-grained LVIS categories in the OOD split (see Appendix A), and suggests the need for stronger semantic generalization or open-vocabulary training strategies. Overall, SFT establishes the interface and ID competence, while IC-GRPO with a difficulty-aware curriculum closes the OOD localization gap.

Comparisons on standard referring-expression benchmarks. Table 5 reports Acc@0.5 on RefCOCO+ (Yu et al., 2016) and RefCOCOg (Mao et al., 2016). The first row (Qwen2.5-VL†) reflects the model's official phrase-grounding pipeline and prompt (Bai et al., 2025a). To standardize interfaces for cross-model comparison, we also evaluate Qwen2.5-VL under a custom prompt (row 2, *); this prompt departs from its native detection-style prompt and substantially lowers accuracy, illustrating the sensitivity of phrase-grounding systems to prompt and decoding design.

Table 6: Localization under the *box-only* setting on RSC in-distribution (RSC-ID) and out-of-distribution (RSC-OOD). Metrics are mIoU and $\mathrm{Acc}@\{0.5, 0.7\}$.

	RSC-ID			RSC-OOD		
Method	mIoU	Acc @0.5	Acc @0.7	mIoU	Acc @0.5	Acc @0.7
Qwen2.5-VL 7B	27.37	24.09	13.42	19.11	13.50	7.70
ScenGround, SFT	51.18	55.80	37.99	40.71	42.65	25.86
ScenGround, GRPO	50.90	54.38	37.01	40.11	40.30	24.50

Under this custom prompt, ScenGround shows two clear trends. First, SFT underperforms on RefCOCO+ but is notably stronger on RefCOCOg, consistent with RefCOCOg's longer, more descriptive phrases being closer to RSC's scenario style. Second, IC-GRPO yields large, consistent improvements across all splits, narrowing the gap to task-specialized pipelines. This suggests that geometry and schema rewards, together with the tag-aware curriculum, help the model resolve intra-class ambiguity and attend to disambiguating cues that typical phrase-grounding prompts would otherwise supply explicitly. Overall, while ScenGround is not optimized for RefCOCO-style prompting, its GRPO stage transfers robustly to these benchmarks, suggesting that curriculum training develops transferable reasoning skills rather than overfitting to a specific prompt template.

Analysis on localization ability. Table 6 isolates localization by asking models to predict a box only, without category prediction. Qwen2.5-VL's box-only performance closely mirrors its full-task localization, suggesting that category prediction is not its primary bottleneck. By contrast, ScenGround delivers strong OOD localization in the box-only regime, substantially outperforming Qwen2.5-VL and confirming that it can ground scenarios from visual cues even when the target category is unseen. The small gap between ScenGround's box-only and full-task results further implies that most remaining OOD degradation comes from naming rather than spatial grounding. Notably, SFT is marginally stronger than GRPO in the box-only setting, while GRPO surpasses SFT on the full task. This suggests that IC-GRPO reallocates capacity to couple localization with scenario understanding rather than optimizing pure box regression.

Analysis on reasoning ability. Table 7 analyzes category prediction under three settings: *Text Only* (scenario s only), *Text-Image* (scenario + image, no

Table 7: Category accuracy on RSC in-distribution (ID) and out-of-distribution (OOD) under three input–output settings. *Text Only*: predict category from scenario only. *Text-Image*: predict category from scenario+image (no box). *Standard*: full task.

	Text Only		Text-Image		Standard	
Method	ID	OOD	ID	OOD	ID	OOD
Qwen2.5-VL 7B	56.24	22.34	77.54	33.33	29.65	19.28
ScenGround-SFT	81.03	10.69	93.66	14.92	92.72	12.30
ScenGround-GRPO	85.13	20.22	94.73	22.68	93.71	20.67

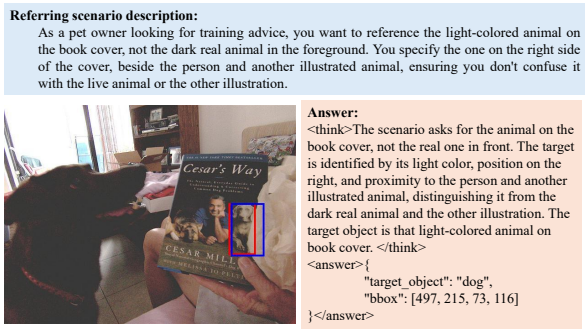


Figure 4: **Qualitative results on RSC.** Red box for ground truth and blue box for ScenGround prediction.

box), and *Standard* (full task). Removing the localization requirement substantially boosts Qwen2.5-VL on RSC-ID but yields a smaller gain on RSC-OOD, showing that its category decisions are sensitive to the localization constraint. By contrast, ScenGround exhibits a small gap between *Text-Image* and *Standard*, indicating that it can jointly localize and classify with minimal degradation. Notably, SFT's strong ID-category accuracy comes at the cost of OOD text-only generalization, while GRPO closes this gap, suggesting that RL improves scenario understanding beyond what schema alignment alone provides. Finally, *Text-Image* outperforms *Text Only* across all models and splits, confirming that visual evidence materially helps resolve scenario ambiguities and that models are not relying solely on language priors.

Qualitative results. Figure 4 illustrates how ScenGround grounds user-driven scenarios. In the shown example, ScenGround correctly distinguishes the *illustrated* animal on the book cover from the real dog by leveraging the cues. *More qualitative examples are in the Appendix G.*

7 Conclusion

We introduce RSC, a benchmark for scenario-based visual grounding where targets are identified from paragraph-length queries specifying user roles, goals, and explicit distractors. RSC provides per-instance difficulty tags, paired reasoning traces, and an OOD split. Evaluations reveal that strong performance on standard referring expression benchmarks does not transfer to scenario-based queries, exposing a systematic failure mode invisible to prior evaluation settings. We also propose ScenGround, a curriculum reasoning model that combines supervised warm-starting with difficulty-aware reinforcement learning, which substantially reduced the localization–semantics trade-off on RSC. We hope RSC provides a useful testbed for studying referential reasoning, with natural extensions toward multi-object, temporal, and interactive grounding.

Limitations and Future Work RSC scenarios are LLM-generated rather than collected from real users. Although our ecological validity study shows that model rankings are preserved across human- and LLM-authored queries for the same instances, the synthetic scenarios exhibit less stylistic diversity than naturalistic language. RSC is also limited to single-target grounding on MS-COCO and LVIS images, which may not generalize to specialized domains or multi-object, video-based, and interactive settings. Future work could complement RSC with human-authored queries to validate ecological validity, extend the benchmark to diverse image domains, and explore multi-object and dialogue-grounded scenario grounding.

Broader Impact RSC is intended to support the development of multimodal models that reason over rich natural language, with applications in embodied assistants and accessibility tools. The benchmark inherits demographic and geographic biases from MS-COCO and LVIS, and LLM-generated scenarios may carry residual hallucinations or culturally specific assumptions despite quality controls. We release annotation logs and metadata to support bias analysis. RSC contains no personally identifiable information, and we do not anticipate direct misuse risks, though we encourage downstream users to consider the implications of scenario-based grounding capabilities in sensitive deployment contexts.

Acknowledgments. This project was partially supported by NSF Award No 2201710, and NSF Award No 2601419, the Ken Kennedy Institute at Rice University, NSF Campus Cyberinfrastructure Award OAC-2019007, and by Rice University's Center for Research Computing (CRC)

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pages 422–440. Springer.

Anthropic. 2025. Claude 3.7. Large language model.

Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, Chun-Fu (Richard) Chen, Alex Bronstein, Kate Saenko, Shimon Ullman, Raja Giryes, Rogerio Feris, and Leonid Karlinsky. 2021. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1801–1812.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025a. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. 2025b. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231*.

Dave Zhenyu Chen and Angel X. Chang. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. 2020. Cops-ref: A new dataset and task on compositional referring expression comprehension.

Zhihong Chen, Ruifei Zhang, Yibing Song, Xiang Wan, and Guanbin Li. 2023b. Advancing visual grounding with scene knowledge: Benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15049.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1769–1779.

Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. 2026. Ref-adv: Exploring mllm visual reasoning in referring expression tasks. *arXiv preprint arXiv: 2602.23898*.

Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, and 1 others. 2025. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. *arXiv preprint arXiv:2504.18904*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364.

Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C. Berg, and Vicente Ordonez. 2024. Improved visual grounding through self-consistent explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13095–13105.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1

others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, and 1 others. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, and 1 others. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017. Referring Expression Generation and Comprehension via Attributes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4866–4874. IEEE.

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

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

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

Navid Rajabi and Jana Kosecka. 2024. GSR-bench: A benchmark for grounded spatial reasoning evaluation via multimodal LLMs. In *NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward*.

Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, and 1 others. 2021. igibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7520–7527. IEEE.

Pengzhan Sun, Junbin Xiao, Tze Ho Elden Tse, Yicong Li, Arjun Akula, and Angela Yao. 2025. Visual intention grounding for egocentric assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2512–2522.

Qwen Team. 2025. Qwen3-vl. Vision-Language Model.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, and 1 others. 2024. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2025. Wizardlm: Empowering large pre-trained language models to follow complex instructions. *URL https://arxiv. org/abs/2304.12244*.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2020. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9952–9961.

Ziyan Yang, Kushal Kafle, Franck Dernoncourt, and Vicente Ordonez. 2023. Improving visual grounding by encouraging consistent gradient-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22915–22924.

Ziyan Yang, Kushal Kafle, Zhe Lin, Scott Cohen, Zhihong Ding, and Vicente Ordonez. 2024. Scord: Subject-conditional relation detection with text-augmented data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5731–5741.

Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2022. PEVL: Position-enhanced pre-training and prompt tuning for vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11104–11117, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan. 2024. Ferret-ui: Grounded mobile ui understanding with multimodal llms. In *European Conference on Computer Vision*, pages 240–255. Springer.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *ECCV*, pages 69–85.

Yufei Zhan, Shurong Zheng, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. 2025. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22947–22957.

Table 8: **Vocabulary size comparison.** Number of unique words across all queries in each dataset's test split.

Dataset	Unique Vocabulary
RefCOCO+	4,240
RefCOCOg	4,917
RSC (Ours)	**9,086**

In this appendix, we provide statistical details of RSC in Section A, RSC data curation details in Section B, RSC data examples in Section C, more details of ScenGround in Section D, tag-level analysis in Section E, linguistic analysis in Section F, ScenGround qualitative examples in Section G, training details in Section H.

A RSC Statistics

Comparisons to RefCOCO+ and RefCOCOg. We compare the query length and instance size distributions. Figure 5 compares query length distributions across RSC, RefCOCO+, and RefCOCOg, illustrating that RSC's scenario-based queries are substantially longer than the short literal phrases used in prior benchmarks. Figure 6 compares instance size distributions, showing that RSC covers a broader range of target scales. Together, these distributions reflect RSC's design goals: longer queries force models to reason over rich contextual descriptions rather than match salient keywords, while the wider size range stresses localization under various controlled conditions. Table 8 further shows that RSC's unique query vocabulary (9,086 tokens) is roughly twice that of RefCOCO+ and RefCOCOg, confirming that the length difference reflects genuine linguistic diversity rather than repetitive padding.

RSC difficulty tag distribution. Figure 7 shows the difficulty tag distributions for RSC-ID and

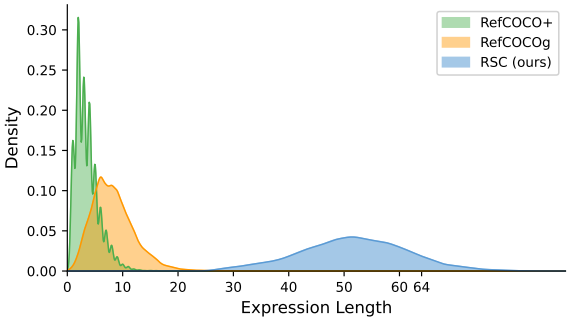


Figure 5: **Query length distribution.** RSC queries are substantially longer than RefCOCO+ and RefCOCOg expressions, reflecting the paragraph-length scenario descriptions that specify user roles, goals, and disambiguating cues. RefCOCO+ and RefCOCOg peak below 10 words; RSC peaks around 50–60 words.

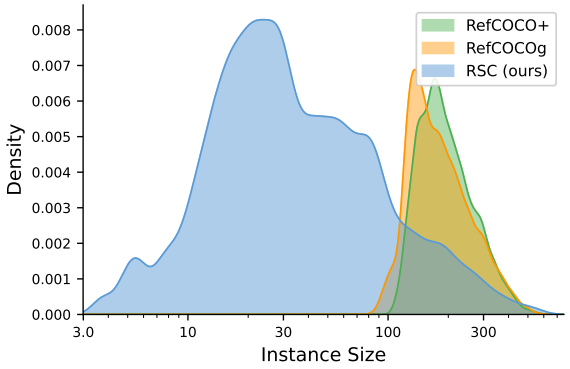


Figure 6: **Instance size distribution.** RSC covers a broader range of instance sizes than RefCOCO+ and RefCOCOg, with notably higher density at smaller instances.

RSC-OOD test splits. The ID split maintains near-balanced marginals across all five axes, reflecting the tag-balanced sampling strategy described in Section 3. The OOD split shows a notable skew toward non-unique (U2) and smaller (S/M) instances, which partially explains the consistently lower accuracy on the OOD split. Distributions across Clutter (C), Overlap (O), and Position (P) remain broadly comparable between ID and OOD, confirming that the primary challenge of the OOD split is category shift rather than a confounding change in scene difficulty along other axes.

RSC scenario linguistic analysis. Figure 8 shows the top-10 words per part-of-speech category across RSC scenario queries. The noun distribution is dominated by general referential terms (*item*, *object*) and scene participants (*man*, *animal*, *photographer*), confirming that scenarios describe roles and contexts rather than naming target categories directly. The verb distribution reflects the action- and goal-oriented nature of scenario queries, with terms such as *distinguish*, *contrast*, *highlight*, and *review* indicating that queries explicitly require disambiguation reasoning. The preposition distribution is rich in spatial terms (*on*, *behind*, *from*, *by*), consistent with RSC's emphasis on relational and positional cues for grounding. Together, these distributions support the claim that RSC queries demand a qualitatively different form of language understanding than the short attribute-and-location phrase queries in standard REC benchmarks.

RSC category distribution. Figure 9 lists the 20 most frequent categories in each test split, and Figures 10 and 11 show the top 80 frequency distributions. RSC-ID covers 79 COCO categories across 4,038 instances, with a moderately long-tailed distribution dominated by common household objects such as *dining table*, *cup*, and *chair*. RSC-OOD draws from 395 LVIS categories across 3,247 instances, with a substantially longer-tailed distribution reflecting LVIS's fine-grained vocabulary, where top categories include clothing items (*jersey*, *jean*, *shirt*) and household furnishings (*cabinet*, *pillow*, *curtain*). Crucially, all OOD categories are disjoint from RSC-ID at both string and synset levels, ensuring that OOD evaluation measures genuine cross-category generalization rather than near-duplicate transfer.

Comparison to natural referential language. RSC's linguistic profile aligns with natural referential tendencies identified in prior work. RefCOCOg, the most naturalistically annotated standard benchmark, has an average query length of 8.4 words and a vocabulary of 4,917 unique tokens; RSC extends this tendency toward longer, richer descriptions (52.7 words, 9,086 tokens), consistent with the shift from object-naming to situation-description that characterizes more complex referential contexts. The dominance of relational verbs (*distinguish*, *contrast*, *highlight*) and spatial prepositions (*behind*, *beside*, *from*) in RSC queries mirrors patterns documented in human referential communication studies, where disambiguating reference in cluttered scenes naturally recruits relational and contextual cues rather than bare category names (Clark and Wilkes-Gibbs, 1986). While RSC scenarios are LLM-generated, the linguistic structure they exhibit reflects these well-documented tendencies rather than purely synthetic artifacts.

To assess ecological validity, we collected

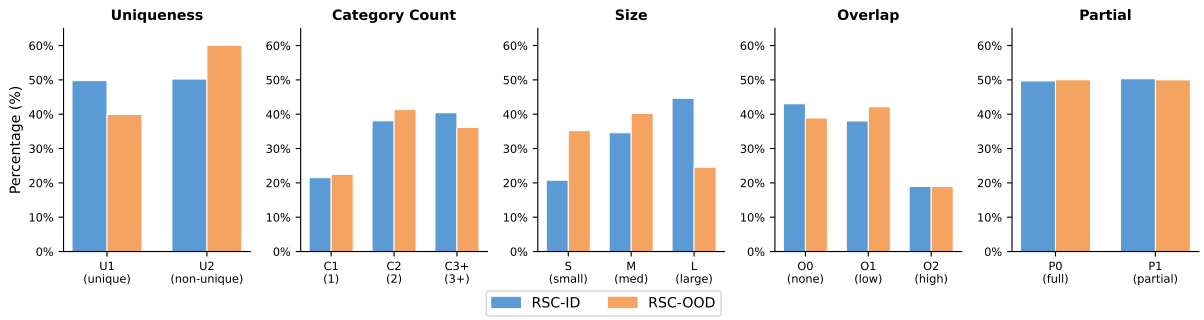


Figure 7: **Difficulty tag distributions for RSC-ID and RSC-OOD.** Each panel shows the per-tag marginal distribution for the in-distribution (RSC-ID) and out-of-distribution (RSC-OOD) test splits across all five axes: Uniqueness (U), Clutter (C), Size (S), Overlap (O), and Position (P).

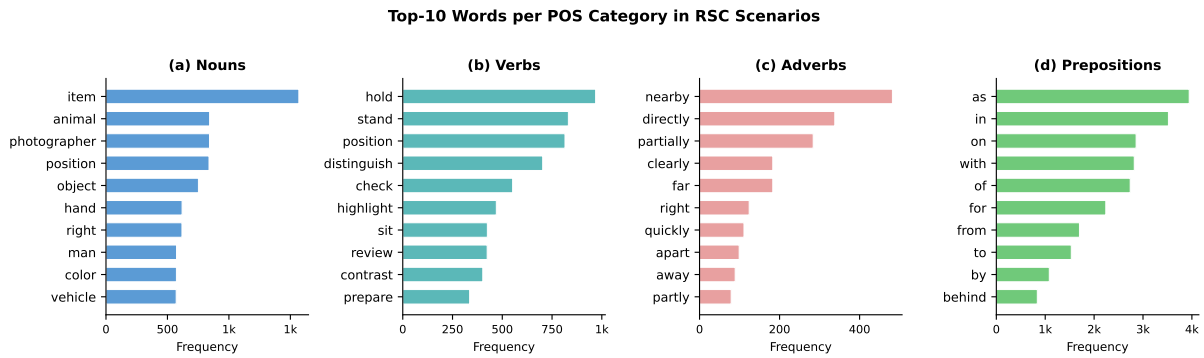


Figure 8: **Top-10 words per part-of-speech category in RSC scenarios.** Frequency counts of the most common nouns, verbs, adverbs, and prepositions across all RSC scenario queries.

Figure 9: **Top 20 most frequent categories** in RSC-ID and RSC-OOD test splits.

Top 20 Most Frequent Categories in RSC

Top 20 Categories RSC-ID (COCO)

Rank	Category	Count
1	dining table	162
2	tie	159
3	cell phone	135
4	bowl	123
5	cup	121
6	knife	119
7	remote	116
8	chair	111
9	fork	104
10	motorcycle	97
11	spoon	96
12	handbag	94
13	bench	92
14	book	92
15	cake	91
16	wine glass	88
17	bottle	88
18	banana	88
19	umbrella	87
20	horse	85

Top 20 Categories RSC-OOD (LVIS)

Rank	Category	Count
1	cabinet	84
2	curtain	83
3	log	79
4	pole	75
5	pillow	72
6	jersey	66
7	blanket	64
8	wheel	61
9	shirt	59
10	trousers	59
11	shoe	57
12	mirror	56
13	jean	53
14	boot	53
15	flag	49
16	drawer	49
17	signboard	48
18	plate	47
19	napkin	47
20	jacket	47

human-authored scenario queries for a subset of 100 RSC instances, asking three annotators to describe the target object without naming its category. Writing a scenario took approximately three minutes per instance, and annotators reported finding the task genuinely challenging: composing a coherent third-person description that provides sufficient disambiguating cues without revealing the target category requires the same kind of intentional reasoning that RSC is designed to evaluate. Comparing human-authored and LLM-generated scenarios for the same targets reveals both differences and commonalities. Human scenarios favor the present tense and exhibit more diverse structural patterns, whereas LLM-generated scenarios follow a more consistent template, which typically opens with a role description, followed by distractor contrasts and attribute cues. Annotators note that synthetic scenarios are less stylistically varied but more detailed and are factually accurate.

To verify that both query types convey equivalent referential content, we asked the three annotators to localize the target from the LLM-generated scenario for the same instances they had authored queries for. All annotators successfully identified the correct target from LLM-generated scenarios, confirming that the referential intent is recoverable from either formulation. Example pairs for the same targets are shown below (Table 10), illustrating that while surface form differs, the underly-

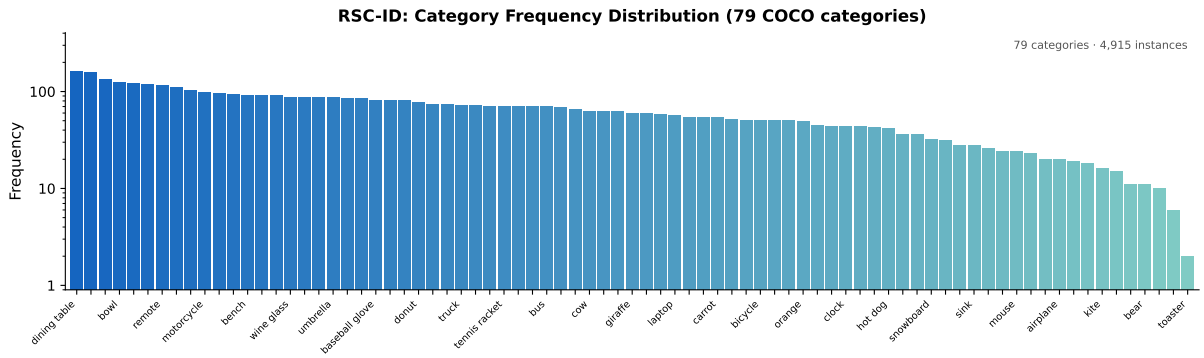


Figure 10: **RSC-ID category frequency distribution.** Frequency of all 79 COCO categories across 4,038 in-distribution test instances, plotted on a log scale. The distribution is moderately long-tailed, with common household objects dominating.

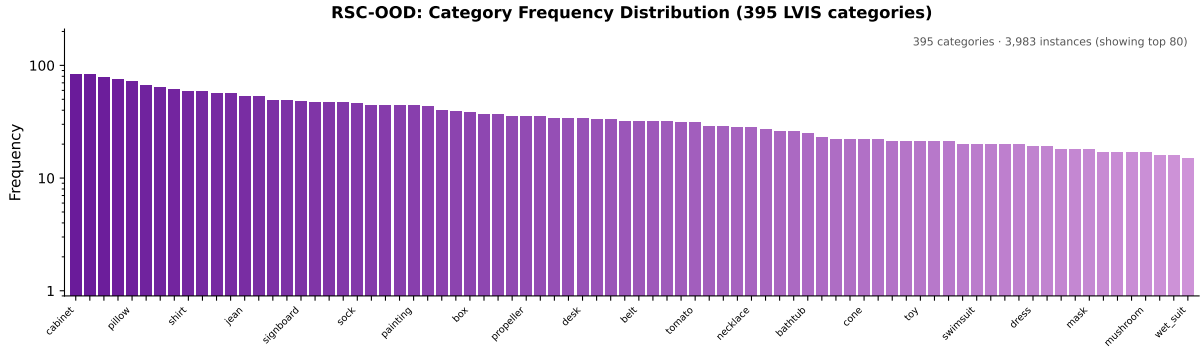


Figure 11: **RSC-OOD category frequency distribution.** Frequency of the top 80 most common categories among 395 LVIS categories across 3,247 out-of-distribution test instances, plotted on a log scale. The long-tailed distribution reflects LVIS's diverse fine-grained vocabulary.

ing disambiguating content is equivalent. Table 9 shows that model rankings are preserved across both query types and absolute performance levels are comparable, supporting the claim that RSC's LLM-generated scenarios probe the same grounding capability as human-authored ones. RSC is designed to test whether models can perform this localization task, as long as the scenario accurately identifies the target through relational and contextual cues, whether human- or LLM-authored.

Table 9: **Model performance on human-authored vs. LLM-generated scenarios** for the same 100 RSC instances. The human-authored scenario query is randomly drawn from one of the three annotators. ScenGround performs comparably on human-authored queries, supporting ecological validity and confirming that training does not overfit to LLM-generated patterns.

	LLM-generated		Human-authored	
Method	mIoU	Cat. Acc	mIoU	Cat. Acc
Qwen2.5-VL	33.4	31.0	31.3	32.0
ScenGround	58.0	96.0	59.2	98.0

B RSC Curation Details

Annotation prompts. Fig. 12 demonstrates the prompt used for RSC annotation generation. For each target, the annotator LLM is shown two views of the *same* image—the full frame and the frame with a red rectangle prior $R(b^*)$ over the target region (the rectangle is *not* part of the object). The prompt supplies context for reasoning only (category hint, image size $H \times W$, pixel box $[x, y, w, h]$, and its normalized form $[x/W, y/H, w/H, h/H]$) and then instructs the model to (i) write one *category-free* referring expression ($\leq$25 words), (ii) write one user-driven *scenario* (3–5 sentences specifying role/goal and $\geq$3 disambiguating cues from attributes/relations/position/affordance), and (iii) produce a brief *thought* explaining how the scenario maps to visual evidence. Outputs must be *JSON-only* with the fixed schema: `object_name`, `acceptable_names` (alias list for internal validation), `bbox_normalized` $[\frac{x}{W}, \frac{y}{H}, \frac{w}{W}, \frac{h}{H}]$, `object_attributes` {`color`, `material`,

Table 10: **Human-authored vs. LLM-generated scenario pairs** for the same targets. Both convey equivalent referential content through different surface forms.

Human-authored	LLM-generated
There is a tourist walking down the street who needs to cross at the intersection. You are looking for a green light on a pole, away from the main intersection light. It's closer to the left side of the street and looks higher than the building.	As a pedestrian waiting to cross, you want to check if it's safe. You look for the green-lit signal on the rightmost pole, above a blue sign, away from the main intersection lights. Its color, position, and separation from the cluster help you identify it.
A parent is watching from the beach, making sure the kid with dark hair is safe. You notice most of his body is submerged, he's close to an adult, and he's the leftmost in the group. Please identify the kid in the crowd.	As a parent watching from the shore, you want to check if your child, who has dark hair and is the leftmost in the group, is safe. They are closest to the water's edge, mostly submerged, and positioned to the left of the shirtless swimmer.
A student is organizing the desk. They are looking for a green-handled item that sits on the desk. There are orange scissors and colorful pens in a mug, but those are not needed. Find the target object to the left of the mug that can be used to cut things.	As a teacher organizing supplies, you need the small green-handled item lying on the desk, not the larger orange-handled ones in the mug. Its bright green color, flat position, and separation from the mug make it easy to spot.

`shape`, `texture`, `size`, `position`, `affordance`, `relation}`, `referring_expression`, `scenario`, and `thought`. The template enforces JSON validity, forbids category leakage in the expression/scenario, requires at least one explicit contrast with plausible distractors, and discourages coordinates in prose in favor of relational cues.

Filtering and quality judge. We first provide the predicted box and the GT bounding boxes from Phase 2 to compute the IoU as the first step of filtering. Then we provide the image, category label, and the generated acceptable name list to GPT-4o (Hurst et al., 2024) for the second step of filtering. To reduce the risk of annotation leakage from the localization prior, we not only explicitly instruct the generator to avoid coordinate-based descriptions in favor of relational cues (e.g., *beside*, *behind*, *closest to*), but also apply a leakage detection to reject any scenario that references the rectangle or uses spatial coordinates verbatim.

Figure 13 shows the prompt used by the quality judge in Phase 3. The system prompt instructs the judge to act as a rigorous visual-grounding quality assessor, determining whether the scenario uniquely and unambiguously identifies exactly one object in the image. The evaluation prompt asks the judge to assess three criteria: UNIQUE (whether exactly one object in the image plausibly matches the scenario), ACCURACY (whether the bounding box correctly and tightly encloses the matching object), and COHERENT (whether the scenario is internally consistent with no contradictory cues). The judge returns a structured JSON response including per-criterion verdicts, a plausible candidate count, a confidence score, and a one-sentence reason. An instance is kept if and only if all three criteria are satisfied (UNIQUE $\wedge$ ACCURACY $\wedge$ COHERENT); borderline cases with low confidence enter the judge refinement loop described in Section 3.

Human audit. Human verification follows the same three-criterion protocol as the quality judge: annotators assess UNIQUE, ACCURACY, and COHERENT for each instance, with a keep decision requiring all three to be satisfied. Three computer science PhD students conducted the audit independently, without knowledge of each other's decisions. To reduce cultural and linguistic bias, annotators were recruited from three different countries (two male, one female). Each annotator reviewed a stratified random sample of 100 instances per audit round. Majority-vote accuracy was 95.6% overall (per-rater: 94%/96%/97%) with great inter-annotator agreement (Fleiss' $\kappa = 0.94$).

C Data Examples

We include additional data examples to illustrate the user-driven scenarios, reasoning traces, and tag structure in RSC. Figure 14 samples from the training split and, for each instance, shows the natural image, the user-driven scenario and the reasoning process. From the held-out test split, Figures 15-16-17 present *easy*, *medium*, and *hard* examples, respectively, reflecting increasing ambiguity (U2), scene clutter (C2/C3), occlusion/overlap (O1/O2), and off-center placement (P1). All examples satisfy the curation gates (alias consistency and IoU threshold; see main text) and omit category names to preserve the scenario.

D ScenGround: Implementation Details

This appendix gives a complete, symbol-level description of ScenGround. All equations correspond one-to-one with our implementation; concrete hyperparameter values appear in Section 5.

```
We provide 2 views of the same image:
Image 1: the full image; Image 2: the SAME image with a RED rectangle around the target region.
Use the RED rectangle only as a localization prior. NOT part of the object.

Context (for reasoning only; do NOT copy literal values in output):
- Category hint: {it.category_name or 'unknown'}  (do NOT name this category in the expression or scenario)
- Image size: {H}x{W}
- Target bbox (pixels [x,y,w,h]): [{int(x)},{int(y)},{int(w)},{int(h)}]
- Target bbox (normalized [x/W,y/H,w/W,h/H]): [{nx},{ny},{nw},{nh}]

Your job:
(A) Write ONE unambiguous referring expression for the target WITHOUT using its category name or synonyms.
(B) Write ONE user-driven scenarios that make the target the natural referent.
(C) Write ONE 'thought' explaining how the object is inferred from the scenario and localized visually.
   - Mention key visual attributes or spatial relations used to connect the description to the image.

Constraints:
- Must use at least 3 cues from: shape, material/texture, color/markings,
  functional affordance, fine-grained part, relative position, relation to distractors, or context/state.
- The RED rectangle is NOT part of the image content.
- Avoid coordinates; use relational descriptions (beside/behind/under/closest to…).
- Include at least one explicit contrast with a plausible distractor visible in the scene.
- Multiple same-category objects likely exist. Please find all relevant objects and be explicit with attributes, relations, and spatial information to disambiguate.
- Expression <= 25 words; natural, concise; no category names.
- Scenario in 3-5 sentences; must NOT name the category. It MUST include:
  (a) user role/situation, (b) goal/motivation, (c) >=3 disambiguating cues (attributes/relations/position).

Please think step by step and output strictly as JSON (no code fences):
{
 "object_name": "<canonical object name within the red box, the red box is NOT part of the object>",
 "accecptable_names": ["<COCO/LVIS category label style synonyms of {it.category_name}>"],
 "bbox_normalized": <[nx,ny,nw,nh]>, the normalized bbox for the annotated object,
 "object_attributes": {
   "color": "<color, the red box is NOT part of the color>",
   "material": "<material>",
   "shape": "<shape>",
   "texture": "<texture>",
   "size": "<size>",
   "position": "<position in image>",
   "affordance": "<functional affordance>",
   "relation": "<relation to other objects>"
 },
 "referring_expression": "<one sentence, <=25 words, no category names>",
 "scenario": "<scenario with user role, goal/motivation, and disambiguating cue>",
 "thought": "<describe how scenario + visual evidence reveal and localize the correct object>"
}
```

Figure 12: The prompt for Referring Scenario Comprehension curation.

D.1 Thought-Primed SFT (TP-SFT)

Objective. Given image x, scenario s, and target text $\mathbf{y}$ (the concatenation of `<think>` and `<answer>` spans), we minimize:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,s,\mathbf{y})\sim\mathcal{D}_{\mathrm{sft}}}\left[\sum_{t=1}^{|\mathbf{y}|}\log p_\theta(y_t \mid x, s, \mathbf{y}_{<t})\right]. \tag{5}$$

Output schema. The instruction requires a single JSON inside `<answer>…</answer>` with two keys: `target_object` (canonical category string) and `bbox` (four pixel integers in xywh, clipped to image bounds). Several key aliases are accepted at scoring time for robustness.

Curriculum. TP-SFT draws from the easy RSC slice (instances with D_i below the easy-percentile threshold δ_{easy}) to stabilize schema and trace learning before RL. The trained checkpoint is saved as the reference policy π_{ref} that anchors KL regularization in Stage 2.

D.2 Incentive-Curriculum GRPO (IC-GRPO)

GRPO objective. For each item (x_i, s_i) we sample K completions $\{c_{i,k}\}_{k=1}^{K} \sim \pi_\theta(\cdot \mid x_i, s_i)$ and

```
You are a rigorous visual-grounding quality assessor.
Your job is to determine whether a given scenario description uniquely and unambiguously identifies exactly ONE object in the image.
You MUST output ONLY a valid JSON object — no markdown fences, no extra text.

Image: <attached>
Category: {category}
Scenario: {scenario}

Evaluate the following:
1. UNIQUE: Can you identify exactly ONE object in the image that matches the referring scenario? Or could two or more objects plausibly match?
2. ACCURACY: Does the bounding box correctly and tightly enclose the object that matches the referring scenario?
3. COHERENT: Does the scenario make internal sense (no contradictory clues)?

Respond with ONLY this JSON (no markdown):
{{
 "unique": true | false,
 "num_plausible_candidates": <integer, how many objects could match>,
 "accuracy": true | false,
 "coherent": true | false,
 "keep": true | false,
 "confidence": <float 0-1, your confidence in keep/reject decision>,
 "reason": "<one sentence explanation>"
}}

Decision rule for "keep":
 - keep=true  iff unique=true AND accuracy=true AND coherent=true
 - keep=false otherwise
```

Figure 13: **Quality judge prompt.** The judge assesses three criteria: uniqueness, bounding box accuracy, and scenario coherence. It applies a strict conjunction rule: an instance is kept only when all three criteria are satisfied.

compute group-relative advantages:

$$A_{i,k} = r_{i,k} - \frac{1}{K}\sum_{k'=1}^{K} r_{i,k'}. \tag{6}$$

The KL-regularized objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_i\left[\frac{1}{K}\sum_{k=1}^{K} A_{i,k}\log\pi_\theta(c_{i,k} \mid x_i, s_i)\right] - \beta\,\mathrm{KL}(\pi_\theta \,\|\, \pi_{\text{ref}}). \tag{7}$$

The KL coefficient β is adapted online by the `AdaptiveKLScheduler`: if the observed KL exceeds the target band $[\kappa_{\text{tgt}} - \kappa_{\text{tol}},\ \kappa_{\text{tgt}} + \kappa_{\text{tol}}]$, β is multiplied by $\mu_\uparrow$; if it falls below, by $\mu_\downarrow$; and β is clipped to $[\beta_{\min}, \beta_{\max}]$.

Box normalization. A predicted 4-vector $\widehat{b} = [x, y, a, b']$ is mapped to a valid box $\widetilde{b} \in \mathbb{Z}^4_{\geq 0}$ by the following priority logic:

1. **Prefer xywh:** if (x, y) lie inside the image and (a, b') are valid width/height that fit within bounds, interpret as (x, y, w, h).
2. **Try xyxy:** else if $a > x$ and $b' > y$ and all values are within image bounds, convert (x_1, y_1, x_2, y_2) to xywh.
3. **Clamp:** otherwise clamp aggressively to image bounds with minimum 1-px side length.

The out-of-bounds indicator is $\mathbb{1}_{\text{OOB}} = \mathbf{1}\{\widehat{b} \neq \widetilde{b}\}$. All IoU values are computed in xyxy after conversion.

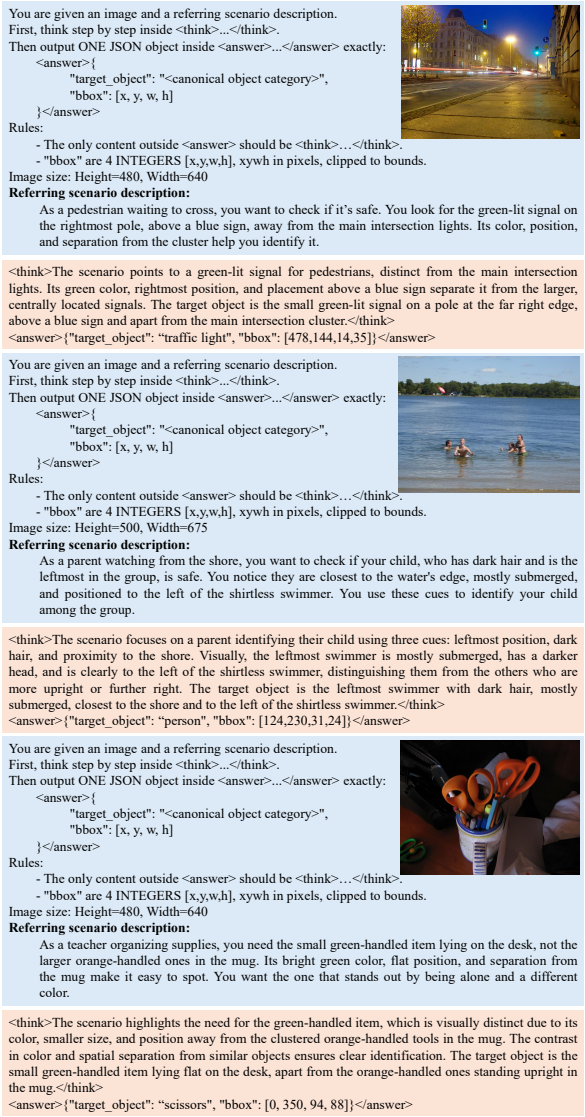


Figure 14: RSC training examples.

Geometry reward. Let $\sigma(z) = (1 + e^{-z})^{-1}$ be the logistic function, $c(\cdot)$ the box center, and $\mathrm{diag} = \sqrt{W^2 + H^2}$ the image diagonal. Define the normalized center distance $d = \|c(\widetilde{b}) - c(b^*)\|_2/\mathrm{diag}$. The geometry reward is:

$$r_{\mathrm{iou}} = \min\Big\{1,\ \underbrace{\mathrm{IoU}(\widetilde{b}, b^*)}_{\text{base}} + \alpha_1\, \sigma\Big(\frac{\mathrm{IoU}(\widetilde{b},b^*)-\tau_1}{\kappa}\Big) + \alpha_2\, \sigma\Big(\frac{\mathrm{IoU}(\widetilde{b},b^*)-\tau_2}{\kappa}\Big) + \alpha_c\, \exp\Big(-\frac{d^2}{2\sigma_c^2}\Big)\Big\} - \alpha_{\mathrm{oob}} \mathbb{1}_{\mathrm{OOB}}, \tag{8}$$

where $\tau_1 < \tau_2$ are the two IoU operating points, κ controls logistic width, α_1, α_2 are the corresponding bonus magnitudes, α_c and σ_c govern the center-consistency term, and α_{oob} is the out-of-bounds penalty.

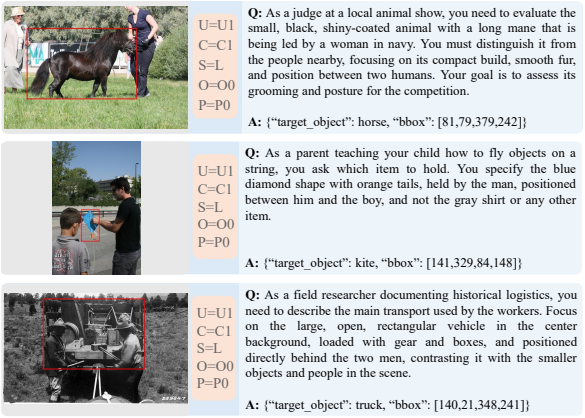


Figure 15: RSC examples with difficulty level "easy".

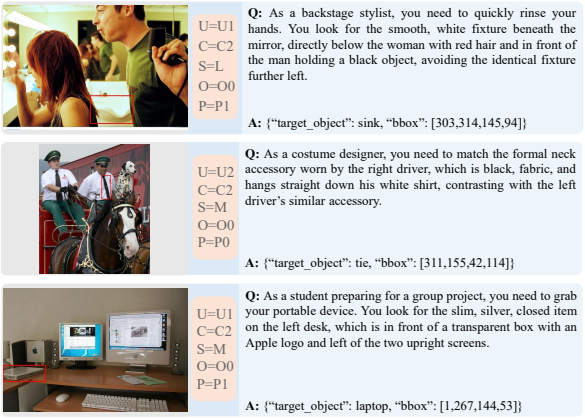


Figure 16: RSC examples with difficulty level "medium".

Category reward. Let $\mathcal{A}^{\mathrm{can}}$ be the canonical name set and $\mathcal{A}^{\mathrm{norm}}$ the normalized alias set. Define an IoU gate:

$$g_{\mathrm{iou}} = \begin{cases} 1, & \mathrm{IoU}(\widetilde{b}, b^*) \geq \tau_g, \\ g, & \mathrm{IoU}(\widetilde{b}, b^*) < \tau_g, \end{cases} \tag{9}$$

where τ_g is the gate threshold and $g < 1$ is the halving factor. With token-level Jaccard similarity $\mathrm{Jac}(\cdot,\cdot)$ computed after lowercasing and light stemming:

$$r_{\mathrm{cat}} = g_{\mathrm{iou}} \times \begin{cases} 1, & \widehat{y} \in \mathcal{A}^{\mathrm{can}}, \\ \eta, & \widehat{y} \in \mathcal{A}^{\mathrm{norm}}, \\ \rho_l + \rho_s \max\limits_{a \in \mathcal{A}^{\mathrm{norm}}} \mathrm{Jac}(\widehat{y}, a), & \text{otherwise}, \end{cases} \tag{10}$$

where η is the alias credit, and ρ_l, ρ_s map Jaccard $\in [0,1]$ to the soft-overlap range $[\rho_l,\ \rho_l + \rho_s]$.

Format and structure rewards. Let $\delta_{\mathrm{tag}} \in \{0,1\}$ indicate the presence of answer tags, $\delta_{\mathrm{json}} \in \{0,1\}$ whether a parseable JSON is found inside, and $\delta_{\mathrm{keys}} \in \{0,1\}$ whether both a bbox key and a

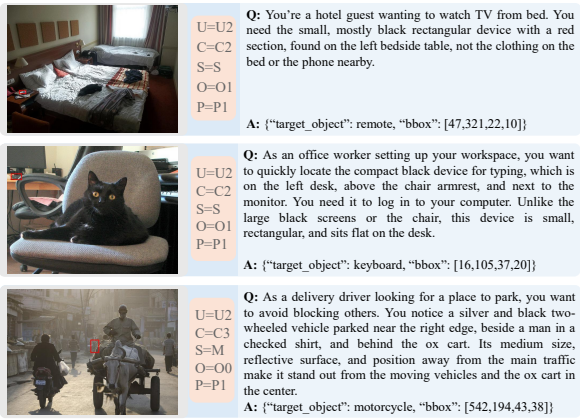


Figure 17: RSC examples with difficulty level "hard".

class key are present:

$$r_{\text{fmt}} = \begin{cases} +1, & \delta_{\text{tag}} = 1 \text{ and } \delta_{\text{json}} = 1, \\ -1, & \text{otherwise}, \end{cases} \tag{11}$$

$$r_{\text{struct}} = \max\{\gamma_{\text{tag}}\, \delta_{\text{tag}} + \gamma_{\text{key}}\, \delta_{\text{keys}},\ \gamma_{\min}\}\,, \tag{12}$$

where γ_{tag}, γ_{key}, and $\gamma_{\min}$ are the tag, key, and floor coefficients.

Total reward and weight annealing. The scalar reward is:

$$r = w_{\text{iou}}\, r_{\text{iou}} + w_{\text{cat}}\, r_{\text{cat}} + w_{\text{fmt}}\, r_{\text{fmt}} + w_{\text{struct}}\, r_{\text{struct}}. \tag{13}$$

Weights are linearly annealed from start to late values over the first p_{anneal} fraction of training:

$$w_h(s) = w_h^{\text{start}} + p(s)\,(w_h^{\text{late}} - w_h^{\text{start}}), \tag{14}$$
$$p(s) = \min\left\{1,\ \frac{s}{p_{\text{anneal}} \cdot S}\right\},$$

where s is the global step and S the total steps.

Tag-aware curriculum. IC-GRPO samples RSC using difficulty scores D_i and tag-based marginals. Let $\Pi^{(m)} = (\pi_{\text{easy}}^{(m)}, \pi_{\text{med}}^{(m)}, \pi_{\text{hard}}^{(m)})$ denote the easy/medium/hard sampling mixture for RL stage m. The mixture shifts between stages to progressively expose the policy to harder U2, C3, O2, and P1 instances, addressing reward sparsity by first establishing reliable IoU signals on easier cases.

Prompt-template ensemble (PTE-8). At each training step we uniformly sample one of T prompt paraphrases. All templates enforce the same output schema; rewards are logged per template for analysis but supervision is identical, improving robustness to surface query variation without changing the learning objective.

Table 11: **Tag-level analysis on RSC-OOD.** mIoU and category accuracy (Cat Acc) broken down by difficulty tag for the baseline (Qwen2.5-VL), SFT, and GRPO.

	mIoU (%)			Cat Acc (%)		
Tag	Baseline	SFT	GRPO	Baseline	SFT	GRPO
U1	27.24	41.71	46.78	26.24	13.57	22.20
U2	16.98	26.44	31.67	16.52	11.71	20.28
C1	22.85	36.06	39.54	12.32	10.53	19.26
C2	21.30	33.00	39.29	19.93	12.22	21.33
C3	20.87	31.44	36.40	27.82	14.31	22.18
S	2.56	14.85	17.18	14.85	12.41	21.87
M	16.50	33.49	38.78	20.56	12.60	21.00
L	49.59	52.64	60.67	27.67	12.57	20.51
O0	18.03	30.11	34.69	20.30	14.81	22.77
O1	21.18	33.53	39.29	19.80	12.08	21.83
O2	29.40	38.73	43.74	24.15	8.91	16.21
P0	16.38	33.01	37.36	21.34	13.96	22.38
P1	26.77	33.39	39.38	20.29	11.08	19.85

E Tag-level Analysis

Table 11 breaks down performance by difficulty tag on RSC-OOD. For mIoU, GRPO consistently improves over SFT across all tags, with the largest relative gains on small instances (S: 2.56→17.18) and high-overlap instances (O2: 29.40→43.74), confirming that the difficulty-aware curriculum meaningfully improves localization on the hardest cases. Large objects (L) achieve the highest absolute mIoU (60.67), as expected from their visual salience. For category accuracy, a notable pattern emerges: SFT drops below the baseline across nearly all tags, indicating that schema alignment reduces OOD category generalisation. GRPO recovers and surpasses the baseline in most cases, demonstrating that RL restores and extends semantic understanding on unseen categories beyond what SFT alone provides. High-overlap instances (O2) remain hard for category naming across all methods, suggesting that visual congestion impairs both localisation and semantic inference simultaneously.

F Linguistic Difficulty Analysis

The U/C/S/O/P tags capture visual grounding difficulty. We complement them with three automatic per-instance linguistic proxies: query length $|q|$, sentence count, and distractor-mention count (explicit contrast phrases such as *not the*, *rather than*, *unlike*).

Table 12 reports Spearman ρ between each proxy and per-instance IoU on RSC-ID. All three corre-

Table 12: Spearman ρ between linguistic proxies and per-instance IoU (RSC-ID). * $p<0.05$, ** $p<0.01$.

Proxy	Qwen2.5-VL	Qwen3-VL	InternVL	ScenGround
$\lvert q\rvert$ (words)	-0.115*	-0.078*	-0.082*	-0.052*
#sentences	-0.074**	-0.030	-0.030	-0.031*
#distractors	-0.115*	-0.093*	-0.078*	-0.048*

Table 13: mIoU by distractor mentions × Uniqueness tag (RSC-ID). †Small cell ($n\approx44$) — interpret with caution.

Model	D=0/U1	D=0/U2	D=1/U1	D=1/U2	D≥2/U1	D≥2/U2
Qwen2.5-VL	36.1	31.8	32.2	23.3	21.3	**17.2**
Qwen3-VL	19.3	11.2	14.3	8.6	8.1	14.7†
InternVL	21.6	11.7	16.7	9.3	12.7	15.9†
ScenGround	**61.7**	48.0	58.3	42.2	51.6	39.0

late negatively and significantly with IoU across all models, confirming that linguistic complexity systematically predicts model failure. Tertile breakdowns confirm monotonic degradation: Qwen2.5-VL drops from 33.16 to 24.74 mIoU (L1→L3 by length) and from 33.45 to 19.51 (D=0→D≥2 by distractor count). The distractor drop is particularly striking: more distractor mentions provide more disambiguating information to a human reader, yet models perform worse, suggesting models are distracted by competing object references rather than using them as cues.

Linguistic and visual difficulty are complementary. Table 13 cross-tabulates distractor-mention count with the Uniqueness tag (U1 = unique category; U2 = non-unique with same-category visual distractors). The joint slice D≥2 ∩ U2 is the hardest combination: Qwen2.5-VL drops from 36.1 (D=0/U1) to 17.2 (D≥2/U2), while ScenGround drops from 61.7 to 39.0 (37%). If the two axes were redundant, the joint slice would not be substantially harder than either alone; the compounding effect confirms they expose distinct failure modes. Notably, ScenGround exhibits the smallest ρ across all proxies (-0.05), indicating that curriculum training improves robustness to linguistic complexity. We release these proxies as additional per-instance fields alongside RSC.

G Qualitative Examples

Figure 18 presents further qualitative examples illustrating how our ScenGround–trained model uses scenario-aware reasoning to improve grounding performance. For each case, we show the input image together with the user scenario, the ground-truth target region, and the model's prediction. The examples span challenging settings such as cluttered scenes, small or partially occluded objects, and pragmatically underspecified descriptions.

H Experimental Setup

Base model. All experiments use Qwen2.5-VL-7B-Instruct (Bai et al., 2025a) as the backbone, with images resized to 512×512.

TP-SFT hyperparameters. AdamW optimizer, learning rate 5×10^{-6}, cosine schedule with warmup ratio 0.15, 5 epochs, gradient accumulation steps 4. The SFT split targets easy-purity $\delta_{\text{easy}} \geq 0.70$.

IC-GRPO hyperparameters. We run two RL stages. Stage 1: AdamW, lr 1×10^{-6}, cosine schedule, warmup ratio 0.05, 5 epochs, K=6 rollouts per prompt, generation batch 240, temperature 0.8, top-p 0.95, max completion length 256. Stage 2: AdamW, lr 2×10^{-6}, cosine schedule, warmup ratio 0.05, 2 epochs, K=12 rollouts per prompt, generation batch 288, temperature 0.9, top-p 0.92, max completion length 160. Both stages use DeepSpeed ZeRO-3, gradient checkpointing, and bf16 precision.

Adaptive KL scheduler. Initial $\beta_0 = 2\times10^{-2}$. Stage 1: κ_{tgt}=0.13, κ_{tol}=0.03, $\mu_\uparrow$=1.5, $\mu_\downarrow$=0.66. Stage 2: κ_{tgt}=0.15, κ_{tol}=0.03, $\mu_\uparrow$=1.6, $\mu_\downarrow$=0.66. Both stages clip $\beta \in [5\times10^{-4},\ 5\times10^{-2}]$.

Geometry reward parameters. IoU operating points τ_1=0.50, τ_2=0.70; logistic width κ=0.03; bonus magnitudes α_1=0.30, α_2=0.50; center coefficient α_c=0.02 with bandwidth σ_c=0.20 (as a fraction of image diagonal); OOB penalty α_{oob}=0.05.

Category reward parameters. IoU gate threshold τ_g=0.30, gate factor g=0.5; alias credit η=0.80; soft-overlap range ρ_l=0.40, ρ_s=0.30 (mapping Jaccard $\in [0,1]$ to $[0.40,\ 0.70]$).

Format and structure reward parameters. γ_{tag}=0.25, γ_{key}=0.75, γ_{min}=-0.50.

Reward weight schedule. Annealing completes by p_{anneal}=0.60 of total training steps. Concrete start and late values are:

Curriculum mixtures. Stage 1 RL uses $\Pi^{(1)}$=(0.70, 0.30, 0.00); Stage 2 RL uses $\Pi^{(2)}$=(0.20, 0.60, 0.20) for (easy, medium, hard) respectively.

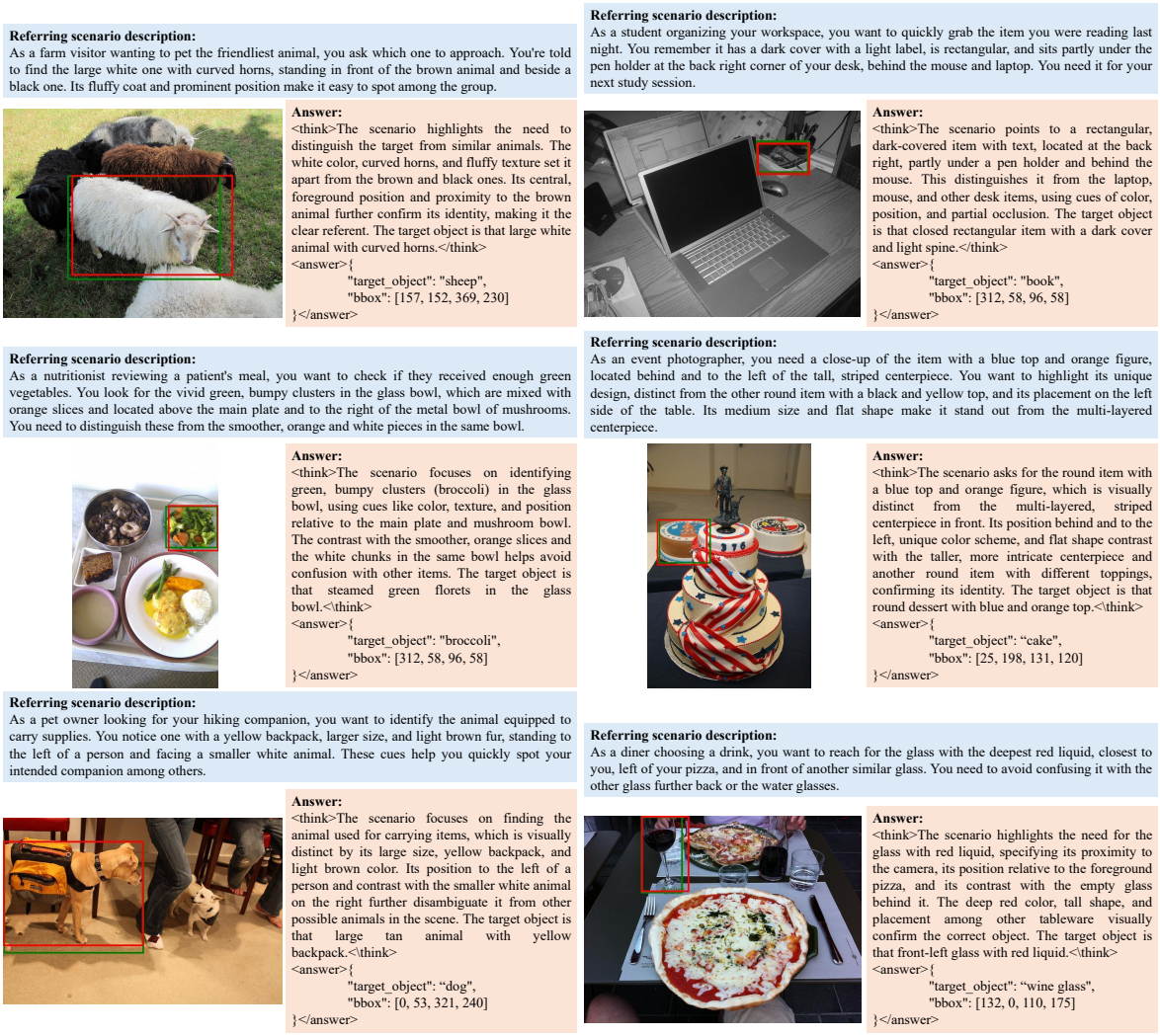


Figure 18: Qualitative examples of ScenGround. Green for the GT box and red for the Predicted box.

Table 14: Reward weight schedules. Stage 2 weights are set by the `RewardWeightScheduler` from step 0, overriding the initialisation values.

	w_{iou}	w_{cat}	w_{fmt}	w_{struct}
Stage 1	0.75	0.15	0.07	0.03
Stage 2 (start)	0.55	0.25	0.12	0.08
Stage 2 (late)	0.75	0.20	0.04	0.01

Prompt-template ensemble. $T{=}8$ paraphrase templates (PTE-8).