From Pixels to Logic: A Perception-Reasoning Decomposition Framework for Open-World Referring Expression Comprehension

Lihong Huang[1], Sheng-hua Zhong[1*], Zhi Zhang[2], Yan Liu[2]

[1]Shenzhen University, College of Computer Science and Software Engineering, Shenzhen, 518060, Guangdong, China
[2]The Hong Kong Polytechnic University, Department of Computing, Hong Kong, 999077, China
huanglihong23@email.szu.edu.cn, csshzhong@szu.edu.cn, zhi271.zhang@connect.polyu.hk, yan.liu@polyu.edu.hk

Abstract

Recent advances in Referring Expression Comprehension (REC) have been largely driven by supervised learning on curated datasets, where each expression is assumed to refer to exactly one known object. However, such assumptions rarely hold in real-world scenarios, where expressions can refer to multiple objects, fail to refer to any, or involve novel categories and complex semantics. These challenges define the task of open-world REC, which demands robust generalization and structured reasoning beyond the scope of traditional REC methods. In this work, we introduce a novel, training-free framework that decouples visual perception from linguistic reasoning to address open-world REC. Our method first transforms the visual scene into a rich textual representation using an open-vocabulary multimodal perception module. It then employs a reasoning language model to interpret the referring expression and perform explicit logical inference over the perceived scene, enabling transparent decision-making and strong generalization in open-world scenarios. Experiments on three standard REC benchmarks as well as two more challenging ones, gRefCOCO and D^3, demonstrate that our framework achieves highly competitive zero-shot performance, often surpassing supervised baselines.

Code — https://github.com/uplihong/PRDF

Introduction

Referring Expression Comprehension (REC) aims to localize a target object in an image based on natural language (Qiao, Deng, and Wu 2020). As a fundamental multimodal task, REC plays a central role in applications such as visual dialogue and human-robot interaction. It requires models to interpret fine-grained linguistic semantics, perceive complex visual content, and align the two modalities through structured reasoning (Xiao et al. 2024).

Supervised learning has driven substantial progress in REC (Yu et al. 2018; Chen et al. 2021; Luo et al. 2023; Wang, Deng, and Jia 2024). These methods achieve strong performance on benchmarks such as RefCOCO (Yu et al. 2016), but are typically trained and evaluated under two implicit constraints: (1) each referring expression corresponds to a single target object, and (2) all referents are drawn from a predefined set of categories.

In real-world scenarios, referring expressions often involve multiple objects ("the two men in red"), no target ("the car" in an empty street), or categories not seen during training. However, existing REC models tend to predict a single bounding box, even when the query implies otherwise (He et al. 2023; Liu, Ding, and Jiang 2023), and struggle to generalize to compositional queries, unseen words, or novel referents (Sadhu, Chen, and Nevatia 2019; Mi et al. 2024; Liu et al. 2024a). Furthermore, the high cost of annotation limits its scalability. These challenges motivate a broader formulation, open-world REC, which supports one, multiple, or zero referents and generalizes beyond training domains.

Two research lines relate to this setting. Generalized REC (GREC) extends traditional REC by explicitly modeling multi-target and no-target cases (He et al. 2023; Dai et al. 2024). However, these methods typically rely on task-specific supervision and complex training pipelines. In contrast, zero-shot REC (Han et al. 2024; Chen and Chen 2025) leverages pretrained vision–language models and avoids finetuning on target datasets, but still focuses almost exclusively on single-target expressions. Conceptually, open-world REC extends zero-shot REC by additionally requiring models to handle variable referent counts and reject absent referents, resulting in a substantially harder task.

To address these challenges, we introduce a training-free, decoupled perception–reasoning framework tailored for open-world REC. Unlike conventional REC, the open-world scenario demands not only flexible perception over unseen categories and varying numbers of referents, but also robust reasoning to interpret ambiguous, compositional, and logically structured expressions. Inspired by the strong semantic generalization and structured-inference capability of recent reasoning-oriented language models (Guo et al. 2025), we explicitly incorporate these strengths into our design. We propose the Perception–Reasoning Decomposition Framework (PRDF), which cleanly separates visual perception from reasoning: a perception module first converts the image into open-vocabulary textual scene descriptions, upon which the reasoning module performs explicit logical inference conditioned on the referring expression. This decoupled design enables a unified solution by leveraging its open-vocabulary perception for novel categories and multi-target

*Corresponding author.

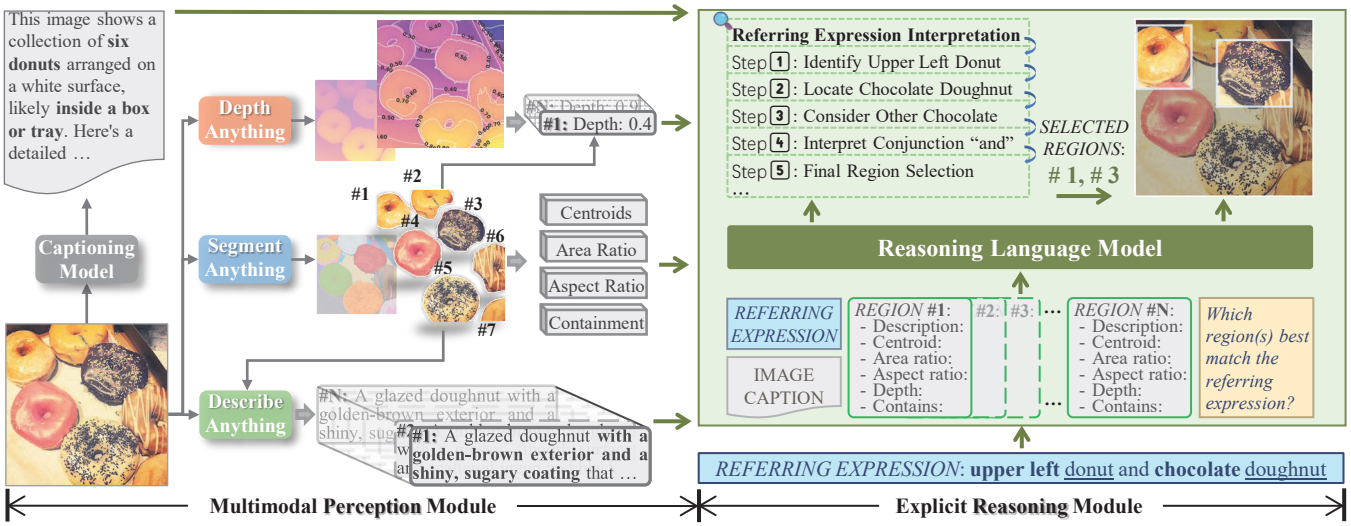

Figure 1: Perception-Reasoning Decomposition Framework.

cases, and its explicit logical reasoning to handle no-target scenarios and diverse language forms.

Our main contributions are as follows: (1) We formalize the task of open-world REC to better reflect real-world scenarios, where referring expressions may involve multiple targets, no targets, or novel object categories. (2) We propose a novel Perception-Reasoning Decomposition Framework (PRDF) that explicitly separates visual perception from linguistic reasoning, enabling decision-making that is modular (components can be independently upgraded), interpretable (producing explicit reasoning traces), and controllable (behavior can be refined via prompts). (3) Extensive experiments on five benchmarks (RefCOCO/+/g, gRefCOCO, D^3) demonstrate that PRDF achieves highly competitive zero-shot performance, outperforming existing zero-shot methods and even surpassing some fully supervised baselines.

Related Work

Referring Expression Comprehension

Classical REC research has progressed from two-stage pipelines (Yu et al. 2018; Chen et al. 2021) to one-stage end-to-end models (Luo et al. 2023) and Transformer-based architectures (Su et al. 2023; Wang, Deng, and Jia 2024), achieving strong performance on standard benchmarks (Yu et al. 2016; Mao et al. 2016). However, these models are fundamentally limited by the assumption that each expression corresponds to a single object from a closed-set vocabulary. Recent efforts have sought to relax these assumptions. Generalized REC (GREC) (He et al. 2023; Dai et al. 2024; Yin et al. 2025) extends the standard formulation to include multi- and no-target expressions. Meanwhile, zero-shot REC (Subramanian et al. 2022; Han et al. 2024; Chen and Chen 2025; Wang et al. 2025) leverages pretrained vision–language models to achieve category generalization without additional annotations.

Together, GREC and zero-shot REC represent important steps toward open-world REC, yet challenges remain in handling ambiguous, compositional, and multi-target expressions under finetuning-free settings.

Open-World Visual Perception with Foundation Models

Open-world visual perception aims to recognize and describe novel objects beyond fixed categories. Recent foundation models support this by shifting toward class-agnostic, promptable systems. Segment Anything (SAM) (Kirillov et al. 2023; Ravi et al. 2024) enables category-free instance segmentation, generating object masks without semantic labels. Describe Anything (DAM) (Lian et al. 2025) augments SAM with vision-language models (Radford et al. 2021; Lin et al. 2024) to produce open-vocabulary descriptions for each region, capturing attributes like color, shape, and material. Depth Anything (Yang et al. 2024a,b) adds monocular depth, enriching spatial context.

Our perception module (Fig. 1, left) integrates these capabilities to produce open-vocabulary scene representations.

Explicit Reasoning with Language Models

General-purpose LLMs often struggle with multi-step logical reasoning (Snell et al. 2024). To address this, reasoning language models (RLMs) aim to emulate System-2 processes via structured inference (Li et al. 2025; Besta et al. 2025). They are trained using: (1) reasoning-oriented prompts like chain-of-thought (CoT) (Wei et al. 2022) and tree-of-thought (Yao et al. 2023); and (2) step-level supervision, including intermediate feedback or rewards (Uesato et al. 2022; Lightman et al. 2023). Recent RLMs such as Deepseek-R1 (Guo et al. 2025) and Qwen-3 (Yang et al. 2025) demonstrate enhanced interpretability and robustness through "show-your-work" training.

We build on these advances by reformulating REC as a language-based reasoning task over perceived scenes.

Method

We introduce the Perception–Reasoning Decomposition Framework (PRDF) for open-world REC. Unlike end-to-end REC/GREC models (Luo et al. 2023; Dai et al. 2024), PRDF explicitly separates perception from reasoning, enabling modularity, interpretability, and zero-shot generalization. The architecture is shown in Fig. 1.

Multimodal Perception Module

The perception module converts an image into an open-vocabulary, attribute-rich scene representation using four pretrained foundation models:

1. Segment Anything. We apply the Segment Anything Model (SAM) (Ravi et al. 2024) to produce a dense set of category-agnostic instance masks $\{M_i\}$, where each mask M_i corresponds to a potential object O_i. This ensures comprehensive object coverage without relying on fixed category vocabularies. From each mask, we compute object-level geometric features f_i^{geom}, including centroid coordinates (x_i, y_i), area ratio a_i, and aspect ratio s_i. These features capture the spatial layout of the scene and facilitate reasoning over position-, size-, and shape-related expressions (e.g., "the tallest person on the left"). Additionally, we derive binary spatial containment relations $r_{i,j}^{\text{contain}} \in \{0, 1\}$ to indicate whether object O_j lies spatially within object O_i, thereby capturing hierarchical spatial structure.
2. Describe Anything. For each region M_i, we employ the Describe Anything Model (DAM) (Lian et al. 2025) to generate a fine-grained textual description d_i^{text} that captures appearance-level attributes such as color, shape, and material. These object-centric descriptions provide rich semantic grounding, while spatial and relational cues are instead modeled by the geometric features f_i^{geom} derived from segmentation.
3. Depth Anything. To complement the 2D geometric layout with coarse 3D structure, we predict the depth z_i of each object O_i using the Depth Anything model (Yang et al. 2024b). This provides an approximate front-to-back ordering of objects, enabling reasoning over expressions involving depth cues (e.g., "person background").
4. Captioning. Finally, we generate an image-level caption C using an off-the-shelf vision-language model. This caption offers high-level contextual semantics that support global scene understanding, help disambiguate vague or ambiguous expressions, and mitigate potential hallucinations in object-level appearance descriptions.

The resulting representation of each object is defined as:

$$O_i = \left\{ d_i^{\text{text}},\ f_i^{\text{geom}},\ r_i^{\text{rel}}, z_i \right\}, \tag{1}$$

where $r_i^{\text{rel}} = \{r_{i,j}^{\text{contain}}\}$ encodes spatial containment relations.

Explicit Reasoning Module

Unlike conventional REC models that rely on end-to-end supervision to implicitly learn grounding patterns, our reasoning module formulates grounding as explicit, query-conditioned inference over object candidates, executed entirely in natural language.

Language-Based Inference. Given the referring expression q, the global caption C, and the structured scene representation $\{O_i\}$, the reasoning language model (RLM) performs language-native inference by consuming a structured natural-language prompt:

$$y = \text{RLM}(\text{Prompt}), \tag{2}$$

where y denotes the predicted referent region(s), accompanied by a trace of intermediate reasoning steps. This formulation enables the framework to perform multi-step deduction, resolve ambiguity, and produce transparent, interpretable outputs, which are typically absent in conventional REC approaches.

Scoring Perspective. The above process can also be viewed as implicitly evaluating how well each object aligns with the query:

$$y = \arg\max_i \ \ \text{score}(q, C, O_i), \tag{3}$$

where the scoring function is instantiated by the RLM's internal reasoning process. In our implementation, we use Deepseek-R1 (Guo et al. 2025), although the framework is compatible with other reasoning-oriented models such as Qwen3 (Yang et al. 2025).

Prompt Construction. To enable transparent and controllable reasoning, we design the prompt in a structured, language-native format. The input includes:

- `REFERRING EXPRESSION:` q
- `IMAGE CAPTION:` C
- `OBJECT DESCRIPTIONS (Total: N):`
  - `Region #i:` d_i^{text}
  - `Centroid:` (x_i, y_i)
  - `AreaRatio:` a_i
  - `AspectRatio:` s_i
  - `Depth:` z_i
  - `Contains: Region #j, ...`
- `OUTPUT INSTRUCTION:` Identify the region(s) that best match the referring expression. Respond in one of the following formats:
  - `NO REGIONS`
  - `SELECTED REGIONS: <number1>, <number2>, ...`

This modular formulation supports explicit reasoning, handles multi-target and no-target expressions, and enables interpretable and controllable decision-making in open-world settings.

Benefits of Perception-Reasoning Decomposition

Separating perception from reasoning offers several key advantages over conventional end-to-end REC models. This design improves interpretability, enhances generalization, and enables flexible deployment in open-world settings.

Method	Setting	RefCOCO			RefCOCO+			RefCOCOg	
		Val	TestA	TestB	Val	TestA	TestB	Val	Test
GPT-4V (Achiam et al. 2023)	Zero-shot	25.5	26.2	24.4	10.6	18.2	8.9	14.3	15.4
CPT (Yao et al. 2024)	Zero-shot	32.2	36.1	30.3	31.9	35.2	28.8	36.7	36.5
Grounding DINO (Liu et al. 2024b)	Zero-shot	50.4	57.2	43.2	51.4	57.6	45.8	67.5	67.1
ReCLIP (Subramanian et al. 2022)	Zero-shot	54.0	58.6	49.5	55.1	60.5	47.4	60.9	61.1
SS-FLAVA (Han et al. 2024)	Zero-shot	52.5	52.7	52.9	50.8	53.4	47.6	61.3	60.9
GPT-4V + DetToolChain (Wu et al. 2024)	Zero-shot	70.0	72.3	67.2	57.3	62.7	52.3	63.5	64.5
EAGR (Bu et al. 2025)	Zero-shot	–	71.0	63.8	–	64.0	<u>53.6</u>	–	71.4
FLORA (Chen and Chen 2025)	Zero-shot	<u>73.7</u>	<u>78.5</u>	**67.8**	<u>63.2</u>	<u>71.6</u>	53.5	<u>72.5</u>	<u>72.1</u>
Ours	Zero-shot	**75.2**	**80.7**	<u>67.7</u>	**65.7**	**71.7**	**57.8**	**75.8**	**75.5**

Table 1: Zero-shot performance on RefCOCO, RefCOCO+, and RefCOCOg benchmarks. All results are reported in terms of IoU@0.5 accuracy (%). Bold denotes the best result; underline indicates the second-best.

- Modularity and Extensibility. Each component in the framework can be independently upgraded or replaced with more advanced models. This modular design facilitates the integration of improvements in vision or language modeling without retraining the entire system.
- Open-World Generalization. Class-agnostic perception ensures coverage of novel object types, while language-based reasoning enables the interpretation of ambiguous, compositional, or under-specified expressions. Together, they support robust performance in settings where categories and referring patterns may not appear in training.
- Training-Free Deployment. All components are based on pretrained foundation models and require no task-specific fine-tuning. This enables zero-shot inference and direct application to new domains without additional annotations.
- Transparency and Interpretability. The system produces explicit intermediate outputs, including structured scene descriptions and reasoning traces. These provide insight into the decision process, support failure analysis, and enable prompt-level correction and refinement.

Experiments

Experimental Setup

Datasets. We evaluate PRDF under the zero-shot setting across five representative benchmarks that collectively cover standard, generalized, and open-world referring scenarios: (1) RefCOCO, RefCOCO+ (Yu et al. 2016), and RefCOCOg (Mao et al. 2016) are widely used REC benchmarks, where each query refers to a single target drawn from a closed set of categories. These datasets primarily test fine-grained alignment between expression and target under standard supervision assumptions. (2) gRefCOCO (He et al. 2023) extends this setting to the open world by introducing more diverse referring expressions, including those with multiple targets, ambiguous references, and queries with no valid target. It thus reflects more realistic and challenging use cases. (3) D^3 (Xie et al. 2023) is a described object detection benchmark designed to evaluate a model's ability to understand attributes, relationships, and their negation. It mainly encompasses two types of queries: presence queries (e.g., "basketball in hand") and absence queries (e.g., "basketball untouched").

Evaluation Metrics. For RefCOCO, RefCOCO+, and RefCOCOg, we report IoU@0.5 Accuracy, which considers a prediction correct if the Intersection over Union (IoU) between the predicted and ground-truth box exceeds 0.5. This is the standard metric for single-target referring expression comprehension (Qiao, Deng, and Wu 2020). For gRefCOCO, we follow the official evaluation protocol and report Precision@(F_1=1, IoU$\geq$0.5) and No-target Accuracy (He et al. 2023). The former assesses both multi-target and no-target cases by requiring all predicted boxes to meet the IoU threshold. The latter measures whether the model correctly predicts that no region corresponds to the given expression when no valid referent exists. For D^3, we report mean Average Precision (mAP) under three settings: *Full*, *Pres* (presence), and *Abs* (absence) (Xie et al. 2023). These reflect the model's ability to reason about whether a described object is present or not, offering a more fine-grained view than traditional object detection metrics.

Implementation Details. Our PRDF framework is constructed using publicly available pretrained foundation models. For segmentation, we use Segment Anything v2.1 (Ravi et al. 2024) with the large Hiera backbone (Ryali et al. 2023). For object-level descriptions, we employ the DAM (Lian et al. 2025); for image-level captions, we use Qwen2.5 (Bai et al. 2025); and for monocular depth estimation, we adopt Depth Anything v2 (Yang et al. 2024b). Experiments are run on V100 (32GB) GPUs using PyTorch 2.6.0.

For reasoning, we adopt Deepseek-R1-0528 (Guo et al. 2025) as the default reasoning language model. Prompts are constructed following the format described in Method Section.

Main Results

Table 1 reports IoU@0.5 on the standard REC benchmarks. PRDF outperforms recent strong baselines, including GPT-4V (Achiam et al. 2023), DetToolChain (Wu et al. 2024), and FLORA (Chen and Chen 2025). These methods typically interleave visual perception with complex reason-

Method	Setting	Val		TestA		TestB	
		Pr.(%)	N-acc.(%)	Pr.(%)	N-acc.(%)	Pr.(%)	N-acc.(%)
MCN (Luo et al. 2020)	Supervised	28.0	30.6	32.3	32.0	26.8	30.3
VLT (Ding et al. 2021)	Supervised	36.6	35.2	40.2	34.1	30.2	32.5
MDETR (Kamath et al. 2021)	Supervised	42.7	36.3	50.0	34.5	36.5	31.0
UNINEXT-R50 (Yan et al. 2023)	Supervised	58.2	50.6	46.4	49.3	42.9	48.2
SimVG (Dai et al. 2024)	Supervised	<u>62.1</u>	<u>54.7</u>	<u>64.6</u>	57.2	**54.8**	57.2
ROD-MLLM (Yin et al. 2025)	Supervised	53.2	51.8	55.4	<u>63.7</u>	51.8	<u>57.9</u>
Ours	**Zero-shot**	**65.4**	**85.2**	**64.9**	**85.8**	<u>52.1</u>	**79.9**

Table 2: Generalized REC results on gRefCOCO. We report Precision@(F_1=1, IoU $\geq$ 0.5) (Pr.) and No-target Accuracy (N-acc.) across val, testA, and testB splits under supervised and zero-shot settings. Bold denotes the best result; underline indicates the second-best.

Method	Full	Pres	Abs
GLIP-T (Li et al. 2022)	19.1	18.3	21.5
Grounding DINO (Liu et al. 2024b)	20.7	20.1	22.5
OFA-DOD (Xie et al. 2023)	21.6	23.7	15.4
FIBER-B (Dou et al. 2022)	22.7	21.5	26.0
InternVL2 (Chen et al. 2024)	25.3	25.7	23.5
ROD-MLLM (Yin et al. 2025)	<u>29.7</u>	<u>30.0</u>	<u>28.7</u>
Ours	**31.6**	**32.2**	**30.3**

Table 3: Mean Average Precision (%) on the D^3 benchmark. We report results for three query types: Full (overall), Pres (presence), and Abs (absence). Bold denotes the best result; underline indicates the second-best.

ing. DetToolChain, for instance, enhances GPT-4V with detection-oriented CoT prompting and visual tools to iteratively refine outputs. FLORA employs an LLM to parse language into a formal structure, which then guides vision-language models like Grounding DINO and CLIP within a probabilistic framework. A key limitation of such interleaved designs is the risk of cross-contamination between perception and reasoning. In contrast, PRDF achieves a new state-of-the-art IoU@0.5 of 75.5% on the RefCOCOg test set among zero-shot methods. This result validates the efficacy of PRDF's strategy, which explicitly decouples perception from reasoning for more robust visual grounding.

Table 2 presents results on gRefCOCO, which evaluates one-target, multi-target, and no-target scenarios. PRDF achieves the highest no-target accuracy (N-acc.) across all splits, significantly outperforming supervised methods such as SimVG (Dai et al. 2024) and ROD-MLLM (Yin et al. 2025). Despite the lack of task-specific supervision, PRDF matches or surpasses their precision, demonstrating strong generalization to open-ended queries.

Table 3 shows results on the D^3 benchmark (Xie et al. 2023). PRDF achieves the highest mAP scores, including 32.2% on presence queries (*Pres*) and 30.3% on absence queries (*Abs*), indicating its ability to reason about both object presence and absence. These results suggest that PRDF can verify the existence of described objects and localize

Figure 2: Qualitative results compared with SimVG.

them based on diverse natural language inputs.

Overall, PRDF demonstrates strong zero-shot performance across all benchmarks, frequently outperforming supervised baselines finetuned on the target dataset. These findings validate the effectiveness of the proposed perception-reasoning decomposition in addressing core challenges of open-world REC.

Qualitative Analysis

We present qualitative comparisons in Fig. 2 and visualize explicit reasoning traces in Fig. 3 to illustrate the effectiveness and interpretability of PRDF.

Fig. 2 highlights PRDF's robustness across diverse refer-

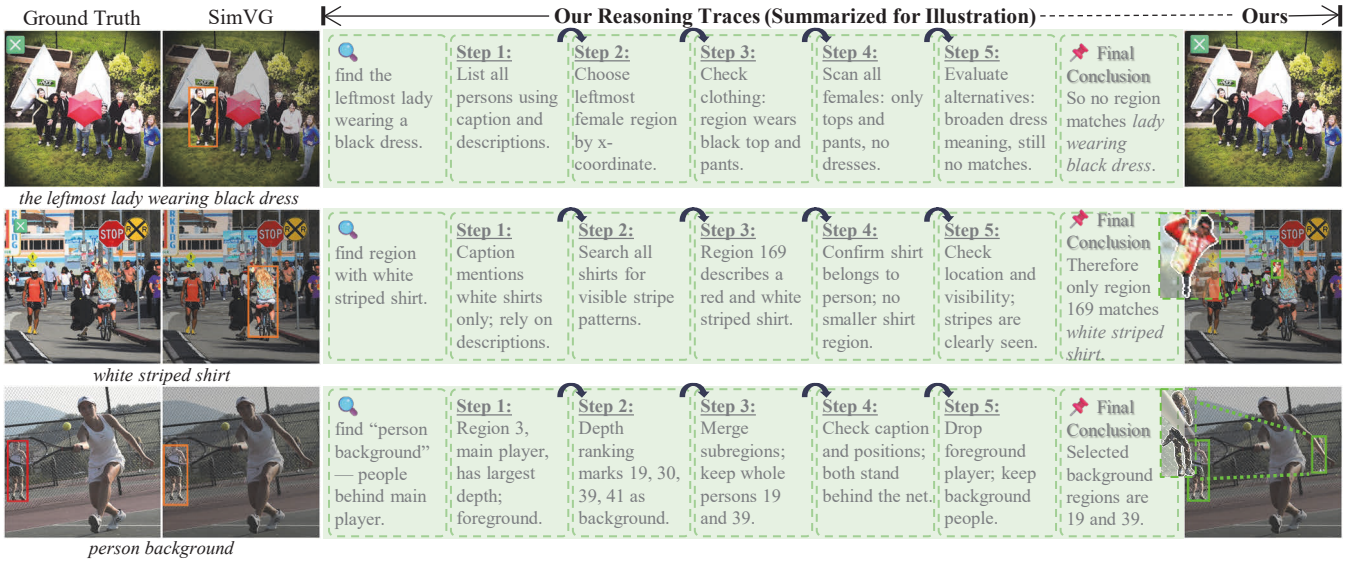

Figure 3: Qualitative results compared with SimVG and some ambiguous cases.

ring scenarios. In the top row, the model accurately grounds two referents using compositional and spatial cues. In the middle row, while both models identify the correct object, PRDF yields a tighter and more precise bounding box. In the bottom row, PRDF correctly distinguishes three adjacent cars in a lineup, whereas SimVG merges two into a single region. These examples demonstrate PRDF's ability to handle multi-target, spatially grounded, and grouped expressions through structured perception and reasoning.

Fig. 3 showcases the transparency afforded by explicit reasoning. In the first example, PRDF correctly resolves a no-target expression by generating a step-by-step reasoning trace that ends with `NO REGIONS`, while SimVG incorrectly returns a bounding box. The second example illustrates a semantic edge case: although no object matches "white striped shirt" exactly, PRDF identifies a small, distant region that partially aligns with the description, reflecting cautious semantic inference. The third example highlights broader coverage: PRDF detects not only the annotated region for "person background" but also an occluded yet semantically consistent figure. Additionally, this example shows that upstream detection errors, such as regions incorrectly described as 'person', are corrected in reasoning steps 2–3 through containment relations.

These qualitative results highlight PRDF's strengths in accuracy, transparency, and generalization, particularly in complex or ambiguous open-world scenarios. Moreover, the reasoning traces reveal that PRDF effectively leverages perceived scene representations for inference, drawing on depth to locate background entities, containment to isolate primary objects, and image-level captions to provide global semantic context.

Component Analysis

We conduct a comprehensive analysis of PRDF to evaluate the contributions of its core components and understand its behavior under varying conditions. This analysis covers four aspects of the framework: (1) the choice of reasoning language model, (2) the impact of different scene representations, (3) the utility of explicit reasoning traces, and (4) the influence of mask quantity on grounding performance.

Language Model	w/ Reasoning	Accuracy(%)
Deepseek-R1-0528	✓	80.7
Qwen3-235B-A22B	✓	78.4
Qwen3-32B	✓	78.9
Deepseek-V3-0324		77.6
Qwen3-235B-A22B		71.4
Qwen3-32B		69.8

Table 4: Ablation study of different language models with and without thinking mode on the RefCOCO TestA.

Ablation of Reasoning Models. We compare different language models in Table 4. Across all variants, enabling explicit chain-of-thought reasoning consistently improves performance, with gains of up to $+7.1\%$. For instance, Qwen3-32B improves from 69.8% to 78.9% when reasoning is enabled. These results demonstrate the benefit of structured reasoning in handling compositional expressions, while also underscoring PRDF's model-agnostic design, which allows flexible substitution with more capable or efficient RLMs.

Ablation of Scene Representations. Table 5 shows the effect of removing individual scene representations from the perception module. Excluding the global scene caption (*w/o caption*) results in the largest performance drop, reducing precision from 52.1% to 42.2% and no-target accuracy from 79.9% to 78.6%, underscoring the importance of high-level contextual cues for resolving referential ambiguity. Removing spatial containment (*w/o containment*) also causes a substantial precision decline (-7.1%), indicating that hierarchi-

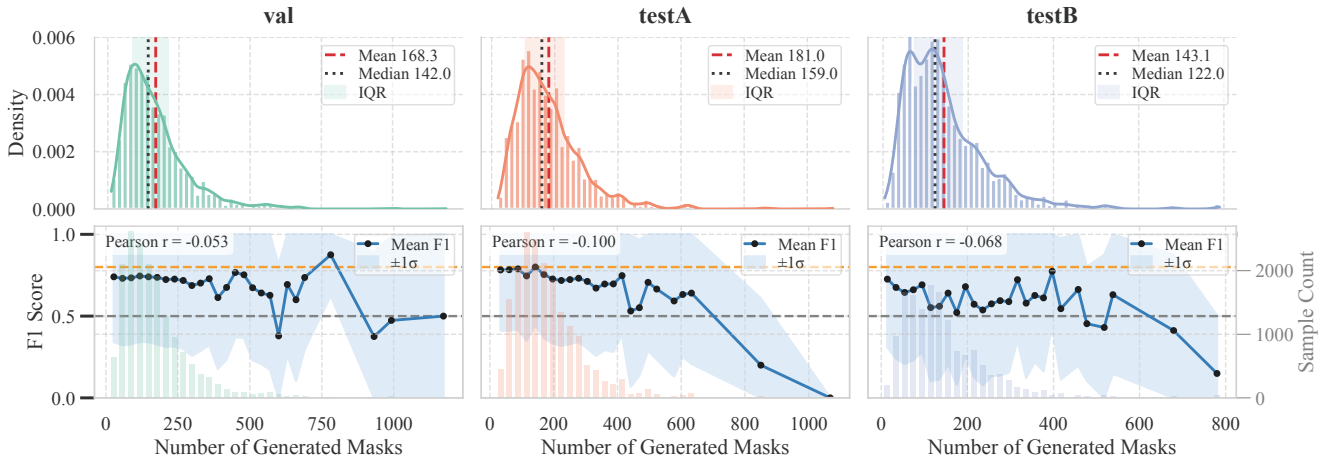

Figure 4: Mask quantity vs. performance. Top: Distribution of the number of generated masks per sample on gRefCOCO's val, testA, and testB splits. Bottom: Corresponding mean F_1 scores binned by mask count.

w/o depth	w/o containment	w/o caption	Pr.(%)	N-acc.(%)
			52.1	79.9
✓			52.1	79.8
	✓		45.0	79.1
		✓	42.2	78.6
✓	✓	✓	37.1	76.9

Table 5: Ablation study of depth, containment, and caption representations in PRDF on the gRefCOCO TestB split.

cal spatial relations aid in identifying primary referents. Removing depth (*w/o depth*) has minimal impact. We attribute this to the nature of the dataset, where most expressions lack fine-grained depth references or rely on coarse distinctions (e.g., "in the background"), for which region size often serves as a sufficient proxy. When all three representations are removed, performance drops further to 37.1% precision and 76.9% no-target accuracy, confirming the complementary role of multimodal cues in supporting robust grounding under open-world conditions.

Effect of Reasoning Traces. A key strength of PRDF lies in its interpretability. By examining reasoning traces, we observe that many errors stem from hallucinated object categories generated by DAM, especially in small or ambiguous regions (e.g., mistaking a tree for an umbrella). To mitigate this, we introduce a lightweight prompt-level refinement: *"Note: Be cautious when interpreting descriptions for objects with very small area ratios, as their category nouns may be prone to hallucination."* This intervention improves performance to 81.3% on RefCOCO TestA and 68.8% on TestB, without any retraining. This result illustrates how explicit reasoning enables modular correction and performance enhancement through natural language, reinforcing PRDF's transparency and controllability.

Statistics of Mask Quantity. We also analyze how the number of segmentation masks affects grounding performance. As shown in Fig. 4, the number of masks follows a long-tailed distribution, occasionally exceeding 1,000 per image. We observe a mild negative correlation between mask count and F_1 score (Pearson r between -0.10 and -0.05), suggesting that excessive segmentation introduces perceptual clutter. This clutter, often caused by scene complexity or over-segmentation, increases reasoning difficulty and can obscure relevant regions, ultimately degrading localization accuracy.

Conclusion and Future Work

We propose PRDF, a novel Perception-Reasoning Decomposition Framework for open-world referring expression comprehension. By decoupling visual perception from language-based reasoning, PRDF transforms raw visual scenes into structured, attribute-rich scene representations and grounds referents using pretrained reasoning language models. This modular and training-free design enables strong zero-shot generalization, supports complex queries such as multi-target and no-target cases, and produces transparent, interpretable predictions.

Extensive evaluations on five benchmarks, including RefCOCO, RefCOCO+, RefCOCOg, gRefCOCO, and D^3, demonstrate that PRDF outperforms both zero-shot and supervised baselines. Analysis results further validate the complementary contributions of structured perception and explicit logical inference. Moreover, PRDF's category-agnostic perception facilitates scalable deployment in open-world settings where exhaustive annotation is impractical.

Despite these advantages, PRDF's performance remains sensitive to the quality of upstream perception and incurs relatively high inference latency due to the use of large reasoning models. Future work includes extending the framework to support multi-turn and multilingual interaction, integrating active perception capabilities, and distilling reasoning behaviors into lightweight, efficient surrogates suitable for real-time use.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (62472291), Natural Science Foundation of Guangdong Province (2025A1515012154), the Science and Technology Innovation Commission of Shenzhen (JCYJ20250604181605008).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Besta, M.; Barth, J.; Schreiber, E.; Kubicek, A.; Catarino, A.; Gerstenberger, R.; Nyczyk, P.; Iff, P.; Li, Y.; Houliston, S.; et al. 2025. Reasoning language models: A blueprint. *arXiv preprint arXiv:2501.11223*.

Bu, Y.; Wu, X.; Cai, Y.; Liu, Q.; Wang, T.; and Huang, Q. 2025. Error-Aware Generative Reasoning for Zero-Shot Visual Grounding. *IEEE Transactions on Multimedia*.

Chen, L.; Ma, W.; Xiao, J.; Zhang, H.; and Chang, S.-F. 2021. Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1036–1044.

Chen, Z.; and Chen, Z. 2025. FLORA: Formal Language Model Enables Robust Training-free Zero-shot Object Referring Analysis. *arXiv preprint arXiv:2501.09887*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. 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*, 24185–24198.

Dai, M.; Yang, L.; Xu, Y.; Feng, Z.; and Yang, W. 2024. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. *Advances in neural information processing systems*, 37: 121670–121698.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16321–16330.

Dou, Z.-Y.; Kamath, A.; Gan, Z.; Zhang, P.; Wang, J.; Li, L.; Liu, Z.; Liu, C.; LeCun, Y.; Peng, N.; et al. 2022. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35: 32942–32956.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Han, Z.; Zhu, F.; Lao, Q.; and Jiang, H. 2024. Zero-shot referring expression comprehension via structural similarity between images and captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14364–14374.

He, S.; Ding, H.; Liu, C.; and Jiang, X. 2023. GREC: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*.

Kamath, A.; Singh, M.; Le, Y.; Misra, I.; Carion, N.; Laurier, N.; Zettlemoyer, L.; Synnaeve, G.; Jégou, H.; and Joulin, A. 2021. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1780–1790.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10965–10975.

Li, Z.-Z.; Zhang, D.; Zhang, M.-L.; Zhang, J.; Liu, Z.; Yao, Y.; Xu, H.; Zheng, J.; Wang, P.-J.; Chen, X.; et al. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Lian, L.; Ding, Y.; Ge, Y.; Liu, S.; Mao, H.; Li, B.; Pavone, M.; Liu, M.-Y.; Darrell, T.; Yala, A.; et al. 2025. Describe anything: Detailed localized image and video captioning. *arXiv preprint arXiv:2504.16072*.

Lightman, H.; Kosaraju, V.; Burda, Y.; Edwards, H.; Baker, B.; Lee, T.; Leike, J.; Schulman, J.; Sutskever, I.; and Cobbe, K. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26689–26699.

Liu, C.; Ding, H.; and Jiang, X. 2023. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23592–23601.

Liu, J.; Yang, X.; Li, W.; and Wang, P. 2024a. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. *arXiv preprint arXiv:2409.14750*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 38–55. Springer.

Luo, G.; Zhou, Y.; Sun, J.; Sun, X.; and Ji, R. 2023. A survivor in the era of large-scale pretraining: an empirical study of one-stage referring expression comprehension. *IEEE Transactions on Multimedia*, 26: 3689–3700.

Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 10034–10043.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Mi, W.; Wang, J.; Zhuang, F.; An, Z.; and Guo, W. 2024. Open-category referring expression comprehension via multi-modal knowledge transfer. *Neurocomputing*, 598: 128063.

Qiao, Y.; Deng, C.; and Wu, Q. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23: 4426–4440.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ryali, C.; Hu, Y.-T.; Bolya, D.; Wei, C.; Fan, H.; Huang, P.-Y.; Aggarwal, V.; Chowdhury, A.; Poursaeed, O.; Hoffman, J.; et al. 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International conference on machine learning*, 29441–29454. PMLR.

Sadhu, A.; Chen, K.; and Nevatia, R. 2019. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4694–4703.

Snell, C.; Lee, J.; Xu, K.; and Kumar, A. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Su, W.; Miao, P.; Dou, H.; Fu, Y.; and Li, X. 2023. Referring expression comprehension using language adaptive inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2357–2365.

Subramanian, S.; Merrill, W.; Darrell, T.; Gardner, M.; Singh, S.; and Rohrbach, A. 2022. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*.

Uesato, J.; Kushman, N.; Kumar, R.; Song, F.; Siegel, N.; Wang, L.; Creswell, A.; Irving, G.; and Higgins, I. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Wang, N.; Deng, J.; and Jia, M. 2024. Cycle-consistency learning for captioning and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5535–5543.

Wang, Y.; Ni, J.; Liu, Y.; Yuan, C.; and Tang, Y. 2025. Iterprime: Zero-shot referring image segmentation with iterative grad-cam refinement and primary word emphasis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8159–8168.

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, Y.; Wang, Y.; Tang, S.; Wu, W.; He, T.; Ouyang, W.; Torr, P.; and Wu, J. 2024. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In *European Conference on Computer Vision*, 164–182. Springer.

Xiao, L.; Yang, X.; Lan, X.; Wang, Y.; and Xu, C. 2024. Towards Visual Grounding: A Survey. *arXiv preprint arXiv:2412.20206*.

Xie, C.; Zhang, Z.; Wu, Y.; Zhu, F.; Zhao, R.; and Liang, S. 2023. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36: 79095–79107.

Yan, B.; Jiang, Y.; Wu, J.; Wang, D.; Luo, P.; Yuan, Z.; and Lu, H. 2023. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15325–15336.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10371–10381.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2024. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5: 30–38.

Yin, H.; Ren, Y.; Yan, K.; Ding, S.; and Hao, Y. 2025. RODMLLM: Towards More Reliable Object Detection in Multimodal Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14358–14368.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. MattNet: Modular Attention Network for Referring Expression Comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1307–1315.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 69–85. Springer.