

New Dataset and Methods for Fine-Grained Compositional Referring Expression Comprehension via Specialist-MLLM Collaboration

Xuzheng Yang, Junzhuo Liu , Peng Wang , Guoqing Wang , Member, IEEE,
Yang Yang , Senior Member, IEEE, and Heng Tao Shen , Fellow, IEEE

Abstract—Referring Expression Comprehension (REC) is a foundational cross-modal task that evaluates the interplay of language understanding, image comprehension, and language-to-image grounding. It serves as an essential testing ground for Multimodal Large Language Models (MLLMs). To advance this field, we introduced a new REC dataset in our previous conference paper, characterized by two key features. First, it is designed with controllable difficulty levels, requiring multi-level fine-grained reasoning across object categories, attributes, and multi-hop relationships. Second, it incorporates negative text and images generated through fine-grained editing and augmentation, explicitly testing a model’s ability to reject scenarios where the target object is absent—an often-overlooked yet critical challenge in existing datasets. In this extended work, we propose two new methods to tackle the challenges of fine-grained REC by combining the strengths of Specialist Models and MLLMs. The first method adaptively assigns simple cases to faster, lightweight models and reserves complex ones for powerful MLLMs, balancing accuracy and efficiency. The second method lets a specialist generate a set of possible object regions, and the MLLM selects the most plausible one using its reasoning ability. These collaborative strategies lead to significant improvements on our dataset and other challenging benchmarks. Our results show that combining specialized and general-purpose models offers a practical path toward solving complex real-world vision-language tasks.

Index Terms—Referring expression comprehension (REC), multimodal large language models (MLLMs), specialist model, reasoning.

I. INTRODUCTION

REFERRING Expression Comprehension (REC) is a fundamental vision-to-language task that aims to localize a target object described by a natural language expression [1]. Successfully addressing this task requires visual understanding, linguistic comprehension, and language-to-image grounding,

Received 27 February 2025; revised 2 May 2025; accepted 1 June 2025. Date of publication 16 June 2025; date of current version 5 September 2025. Recommended for acceptance by N. Sebe. (Xuzheng Yang and Junzhuo Liu are contributed equally to this work.) (Corresponding author: Peng Wang.)

Xuzheng Yang, Junzhuo Liu, Peng Wang, Guoqing Wang, and Yang Yang are with the University of Electronic Science, and Technology of China, Chengdu 610054, China (e-mail: p.wang6@hotmail.com).

Heng Tao Shen is with the University of Electronic Science, and Technology of China, Chengdu 610054, China, and also with Tongji University, Shanghai 200070, China.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TPAMI.2025.3580090>, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2025.3580090

making it an ideal testbed for Multimodal Large Language Models (MLLMs). REC has broad applications in assistive AI systems, such as a robotic assistant identifying “the red cup next to the coffee make” in a cluttered kitchen or an AI-powered medical tool pinpointing a specific lesion in radiology scans based on textual descriptions from doctors.

To advance Referring Expression Comprehension research and its real-world applications, datasets tailored for Multimodal Large Language Models (MLLMs) are essential. Traditional REC benchmarks such as RefCOCO, RefCOCO+, and RefCOCOG [2], [3], [4] lack considerations for compositional reasoning, allowing models to perform well without truly understanding linguistic structure—or in some cases, even without relying on the expression itself [5], [6]. Moreover, most existing datasets do not include negative samples, where the target object is absent from the image. This omission is critical, as real-world REC scenarios often require models to reject incorrect object proposals when the described entity is not present. The lack of such negative examples limits the robustness of current REC models, making them prone to false positives.

To address these limitations in existing datasets, we introduced the FineCops-Ref benchmark in our previous work [7]. FineCops-Ref is specifically designed to enhance fine-grained reasoning in REC by incorporating controlled difficulty levels and meticulously constructed negative samples. The difficulty levels compel MLLMs to reason across object categories, attributes, and multi-hop relationships, categorized based on the number of attributes and relationships necessary for accurate localization. The inclusion of negative samples evaluates model resilience against misalignments and hallucinations, providing a more rigorous assessment of their true visual grounding capabilities. By addressing both compositional reasoning and negative sample inclusion, FineCops-Ref establishes a more realistic and challenging benchmark for advancing REC research in the MLLM era.

Extensive evaluations on the FineCops-Ref benchmark also reveal a critical gap in current vision-language models [7]: while MLLMs demonstrate strong generalization and reasoning abilities, they often struggle with fine-grained localization and compositional grounding, particularly when precise object differentiation is required. In contrast, Specialist Models, which are typically lightweight and trained for specific vision

tasks, excel in efficient and accurate low-level perception but lack the broad contextual reasoning capabilities of MLLMs. Given their complementary strengths, an effective solution is to leverage both models strategically—assigning simpler, perception-driven tasks to Specialist Models while reserving MLLMs for complex reasoning. This collaborative approach not only enhances accuracy and robustness but may also improve computational efficiency, as MLLMs are invoked only when necessary.

Building on this insight, this extended work introduces two collaboration strategies between Specialist Models and MLLMs to improve REC performance. The first strategy, Slow-Fast Adaptation (SFA), introduces an adaptive task routing mechanism that dynamically assigns simpler, detection-oriented tasks to lightweight Specialist Models, while delegating complex tasks requiring compositional reasoning to MLLMs. This adaptive and flexible collaboration not only improves efficiency but, more importantly, enhances overall REC performance. Additionally, we identify common error patterns in both types of models, particularly their tendency to mistakenly identify objects not belonging to the described category. To mitigate this issue, we propose a simple yet effective target-refocus strategy, which obviously reduces localization confusion for both Specialist Models and MLLMs. Notably, the SFA strategy is entirely training-free, making it highly practical for real-world applications. The second strategy, Candidate Region Selection (CRS), aims to improve recall and precision by generating multiple object candidates using Specialist Models, followed by leveraging MLLMs’ reasoning capabilities to select the correct target. While MLLMs inherently possess reasoning capacity, they are not explicitly trained to approach selection tasks as generative problems, limiting their effectiveness in such scenarios. To address this, we conduct instruction tuning on external datasets, ensuring that MLLMs can effectively address selection tasks while avoiding knowledge leakage. This instruction tuning process generalizes well across various REC datasets.

Experiments on three compositional REC datasets demonstrate that our Specialist-MLLM collaboration strategies significantly enhance the performance of both existing Specialist Models and MLLMs, highlighting the generalizability of our approach. Beyond achieving superior performance, we aim for this work to provide valuable insights into solving complex real-world tasks by strategically leveraging existing models for maximum effectiveness, rather than reinventing them. This aligns with the broader trend of rapidly evolving LLM-based agents, emphasizing the importance of optimizing their potential to tackle challenging tasks efficiently.

This work extends our previous conference paper [7], which introduced the FineCops-Ref benchmark. In this journal version, we make three major contributions: (1) we propose two novel Specialist-MLLM collaboration strategies tailored for fine-grained compositional REC; (2) we conduct more extensive experiments across FineCops-Ref and multiple external benchmarks; and (3) we provide deeper analyses and ablations, offering a more self-contained and comprehensive exploration of the task.

II. RELATED WORKS

A. Benchmarking Referring Expression Comprehension

Referring Expression Comprehension (REC) is a vision-language task where models must localize a target object in an image based on a natural language referring expression, requiring compositional understanding of attributes (e.g., color, size), spatial relationships, and object categories to disambiguate the referent from other entities. While early benchmarks like RefCOCO/+g [2], [3], [4] established foundational evaluation protocols, subsequent analyses revealed critical limitations. For instance, [5] and [6] demonstrated that models could exploit dataset biases rather than genuinely parse linguistic structure: [6] found that up to 83.7% of RefCOCOg test instances could be solved via keyword matching or visual feature dominance, bypassing compositional reasoning. To address this, [6] introduced Ref-Adv, a dataset where perturbed expressions refer to alternate objects, forcing models to engage with linguistic structure. Subsequent efforts have prioritized compositional reasoning. CLEVR-Ref+ [8] is a synthetic dataset emphasizing relationships, attributes, and linguistic logic. Cops-Ref [9] and Ref-Reasoning [10] use GQA scene graphs [11] and rule-based methods to create large-scale compositional referring expression comprehension datasets in real-world scenarios. Cops-Ref additionally introduces distracting images based on attributes, relationships, and target names. Recent benchmarks like GITM-MR [12] explores mismatched relationship in the REC task, while ARPGrounding [13] evaluates the model’s compositional reasoning ability by constructing referring expression pairs that target distinct objects of the same category within a single image, differentiated solely through their attributes or relational contexts. RefEgo [14] and OmniLabel [15] consider out-of-distribution scenarios where the referred targets do not exist in the image.

The challenges in rigorously evaluating compositional reasoning are not unique to REC. Across various related vision-language tasks, such as image captioning, image-to-text retrieval, and visual question answering, recent evaluations similarly reveal that current vision-Language Models struggle with compositional understanding [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27]. For instance, prominent benchmarks like FOIL-it [16], VALSE [17], and COCO-Counterfactuals [20] are specifically designed to test compositional understanding by creating “hard negative” image-caption pairs that highlight subtle distinctions. These broader evaluation efforts underscore the critical need for benchmarks that can truly test deep linguistic-visual understanding and resilience to subtle distractors in complex, compositional scenarios – a requirement particularly pronounced in fine-grained REC where minor linguistic cues can drastically change the target.

Building on these advances, we introduced FineCops-Ref in our previous work [7], a new REC benchmark specifically designed to more comprehensively assess the fine-grained compositional reasoning abilities of multimodal models. Based on this challenging benchmark, we propose novel Specialist-MLLM collaboration strategies and provides more extensive experiments to further advance fine-grained compositional REC.

B. Specialist REC Models

Early Specialist Models [28], [29], [30], [31], [32], [33], [34], [35], [36] typically employed pre-trained detectors to generate proposals and locate targets through text-to-region matching. Subsequent approaches transitioned to ViT-based architectures, typically adopting one-stage designs [37], [38], [39], [40]. Recent advancements [41], [42], [43], [44] have shifted towards DETR-based [45] frameworks, utilizing encoder-decoder architectures for multimodal fusion and generating target bounding boxes through object queries. Specifically, MDETR [41] pioneered the integration of free-form text with DETR, developing an end-to-end modulated detector capable of identifying objects in images based on raw text queries. UNINEXT [42] implemented Deformable DETR [46], proposing a universal instance perception model. Grounding-DINO [43], building upon the advanced DINO framework [47], enhanced DINO by incorporating vision-language modality fusion at multiple stages. [48] explored the utilization of hard negative samples during training. More recent models, such as OneRef [49] and SimVG [50] have streamlined the architecture by eliminating complex integration designs and redundant parameters through the implementation of modality-shared feature spaces, employing a unified multimodal encoder. Specialist REC models, typically lightweight and specifically trained for vision tasks, excel in efficient and accurate low-level perception. However, they generally lack broad contextual reasoning capabilities, which limits their performance on complex REC tasks.

C. Multimodal Large Language Models for REC

Unlike Specialist REC models, which usually treat REC as a bounding box regression task, MLLMs often formulate the bounding box prediction as a text generation problem, outputting bounding box coordinates in an autoregressive manner. Recent MLLMs [51], [52], [53], [54], [55], [56], [57], [58] have leveraged the powerful capabilities of LLMs through zero-shot or instruction tuning methods to address complex REC tasks. These MLLMs also known as grounding MLLMs [1]. Shikra [51] is an early model to support region-level input and understanding. GroundingGPT [53] employs a coarse-to-fine training strategy to enhance model's semantic awareness. CogVLM [55] incorporated a visual expert module in each transformer layer to enable dual fusion between vision and language features. CogCoM [56] introduced Chain of Manipulations, encouraging MLLMs to generate responses through evidential visual reasoning. Regarding visual modules, some methods, such as GLaMM [59] and LLaVA-Grounding [60], integrate additional visual components, while others, including Groma [61], VisualCoT [62], Ferret [52], and CoVLM [63], extract regional features as supplementary inputs. Recently, generalist MLLMs, such as Qwen2-VL [64] and InternVL 2.5 [65], have demonstrated significant advancements in grounding capabilities while maintaining strong reasoning performance, achieved through enhanced model architectures, carefully curated datasets, and optimized training strategies.

MLLMs demonstrate superior generalization and reasoning capabilities compared to Specialist Models, making them

particularly well-suited for complex REC tasks demanding deeper contextual understanding. However, MLLMs inherit significant computational burdens from two key aspects. First, they rely on LLMs which possess substantially more parameters than Specialist Models. Second, the attention mechanisms used by MLLMs exhibit quadratic scaling of computational complexity. This scaling is further exacerbated by the increased input length resulting from the integration of numerous visual tokens (e.g., image patches) with textual tokens.

III. FINECOPS-REF BENCHMARK

In this section, we provide a summary of the FineCops-Ref dataset introduced in our previous work [7]. FineCops-Ref is a benchmark specifically designed to address the limitations of existing datasets in evaluating fine-grained compositional reasoning and robustness to negative samples in Referring Expression Comprehension. It comprises both positive and negative samples, carefully constructed to provide a more challenging and realistic evaluation setting for Multimodal models. The dataset construction pipeline is illustrated in Fig. 1. The detailed methodology for constructing FineCops-Ref, including path generation, data categorization, expression generation, human filtering, and the generation of negative data (both expressions and images), was introduced in our previous conference work [7].

A. Key Characteristics

FineCops-Ref is distinguished by the following key characteristics, which make it a suitable and challenging benchmark for evaluating advanced REC models and our proposed collaboration strategies:

Controlled Difficulty Levels: The dataset incorporates three controlled difficulty levels (Level 1, 2, and 3) based on the complexity of fine-grained reasoning required to uniquely identify the target object. This necessitates models to reason across object categories, attributes, and multi-hop relationships. Level 1 involves simple localization without distractors of the same category, Level 2 requires distinguishing the target using a single attribute or relation among same-category distractors, and Level 3 demands the interpretation of two or more relationships and attribute dependencies.

Comprehensive Negative Samples: FineCops-Ref includes both negative expressions and negative images. These negative samples are specifically designed as “hard negatives” with subtle differences from positive instances, evaluating the model’s ability to correctly reject instances where the described object is absent or the expression/image is subtly mismatched. The construction involves various types of challenging negative samples generated through careful editing and perturbation.

B. Dataset Statistics

FineCops-Ref consists of 9,605 positive expressions, 9,814 negative expressions, and 8,507 negative images in test set. Table I provides a comparison between FineCops-Ref and other

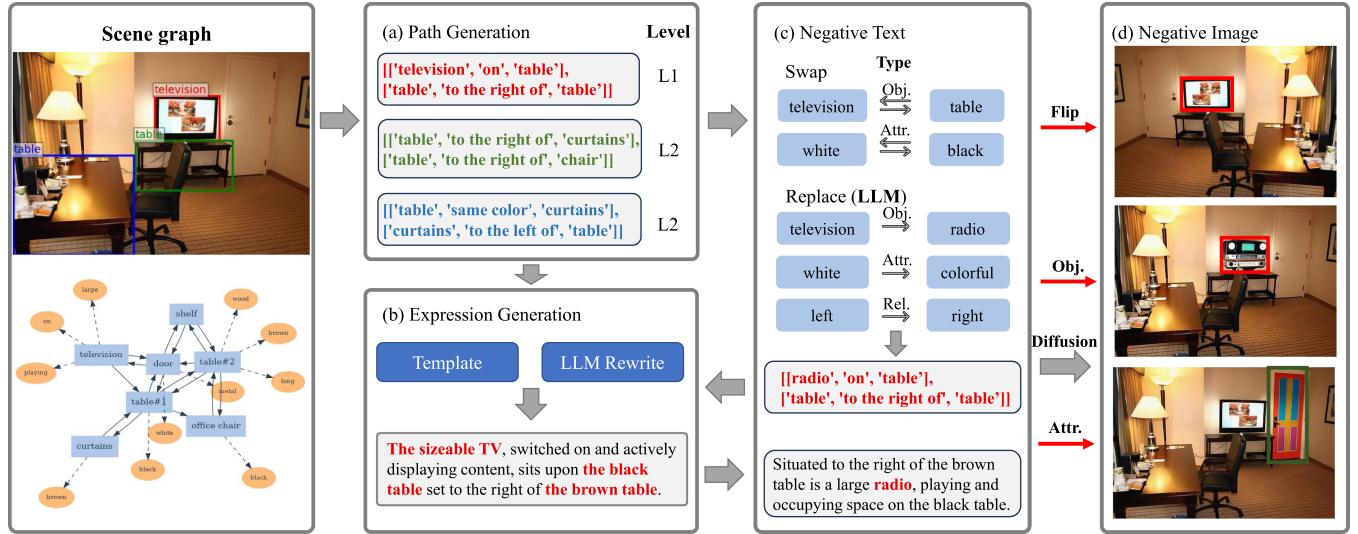


Fig. 1. The data construction pipeline of FineCops-Ref. [7]. Given an image, we first generate paths based on its scene graph. Then, we fill paths into templates and obtain the positive referring expression through LLM rewriting. Meanwhile, we utilize LLM to generate negative expressions, and based on this, we employ diffusion model to create fine-grained editing negative images.

TABLE I
COMPARISON BETWEEN THE PROPOSED BENCHMARK AND OTHER REC BENCHMARKS

Benchmark	Unconstrained	Cops.	Difficulty level	Neg. text	Neg. image	Positive		Negative	
						Expression	Expression	Image	
RefCOCO [3]	✓					10,752	-	-	
RefCOCO+ [3]	✓					10,615	-	-	
RefCOCOG [4]	✓					9,602	-	-	
Ref-Reasoning [10]		✓	✓			34,609	-	-	
Cops-ref [9]		✓			✓	12,586	-	37,758	
Ref-Adv [6]	✓	✓	✓	✓	✓	9,602	3,704	-	
Ours	✓	✓	✓	✓	✓	9,605	9,814	8,507	

Unconstrained indicates the final expression is not constrained by the templates. Cops. indicates compositional reasoning. On the right hand side, the test set count of each benchmark is listed.

TABLE II
DATASET STATISTICS

Set	Positive	Negative expression	Negative image
Train	163,792	80,451	-
Val	18,455	9,029	-
Test	9,605	9,814	8,507

visual grounding benchmarks. FineCops-Ref combines the advantages of unconstrained expression, fine-grained compositional reasoning, difficulty level, and hard negatives at both textual and visual levels. Additionally, we partition the training set and validation set simultaneously as in Table II. For more statistics, please refer to the Appendix, available online.

C. Metrics

To evaluate performance on positive data, we use the common metric Precision@k. When both positive and negative data are present in the test set, we treat the negative samples as distractors for the positive samples, and introduce two additional metrics:

Recall@k: We treat the REC task as a bounding box retrieval problem. For each negative sample paired with its corresponding positive sample, we first obtain the predicted bounding boxes from the model, along with their confidence scores for both positive and negative samples. These bounding boxes are then ranked based on their confidence scores. Recall@k measures the proportion of negative-positive pairs where at least one of the top k predicted bounding boxes has an IoU greater than 0.5 with the ground truth bounding box. It specifically assesses the model's ability to avoid assigning high confidence scores to negative samples. Formally, Recall@k is defined as:

$$\text{Recall}@k = \frac{1}{N} \sum_{i=1}^N \mathbb{1} \left(\max_{j \in \{1, \dots, k\}} \text{IoU}_{i,j} > 0.5 \right), \quad (1)$$

where N represents the total number of negative-positive pairs, and $\mathbb{1}(\cdot)$ is an indicator function that equals 1 if the condition inside is true and 0 otherwise. The term $\text{IoU}_{i,j}$ refers to the overlap between the j th predicted bounding box (ranked based on the confidence scores) and the ground truth bounding box for

the i th pair. Note that for negative samples, there is no ground truth box, meaning the IoU is 0.

Recall@ k is commonly used in retrieval tasks to assess prediction accuracy in the presence of challenging negative samples. Ideally, the model should assign lower confidence scores to negative samples. In our study, we primarily report Recall@1. If the model consistently assigns lower confidence scores to negative samples compared to positive ones, Recall@1 should equal Precision@1.

AUROC: While Recall@ k evaluates how well the model ranks individual negative samples relative to their corresponding positive samples, it does not offer a holistic view of confidence across the dataset. To address this, we use AUROC to measure the overall ability of the model to distinguish between positive and negative samples. AUROC measures the model's ability to correctly rank positive samples higher than negative ones across the datasets, providing a holistic view of its discriminative power.

By combining Recall@ k and AUROC, we ensure a comprehensive evaluation of the model's performance in distinguishing between positive and negative samples in REC tasks. This dual approach addresses both specific ranking and overall confidence.

IV. METHOD

Referring Expression Comprehension presents a unique challenge in computer vision, demanding not only precise object detection but also compositional visual reasoning to ground natural language descriptions to specific image regions. Specialist Models, such as MDETR [41] and Grounding DINO [66], demonstrate efficiency and excel at accurate object detection and grounding. Their strengths lie in their optimized architectures and training on extensive datasets with bounding box annotations, enabling precise box regression and rapid processing. These models deliver exceptional grounding accuracy and high Recall@ k , ensuring that the correct object is often among the top predictions. Indeed, for simpler REC instances where target identification relies primarily on object category, Specialist Models often suffice. However, Specialist Models exhibit limitations in handling the inherent complexity of natural language. They often struggle with referring expressions that require compositional reasoning, disambiguation among multiple objects of the same category.

In contrast, MLLMs exhibit strong visual reasoning capabilities, making them well-suited for complex REC tasks that require deeper semantic understanding. However, they are computationally demanding and not inherently optimized for grounding tasks, as they generate bounding box coordinates directly as part of a text generation process, rather than leveraging intermediate bounding box proposals as informative priors, as specialist models do.

This motivates our collaborative approach, which aims to synergistically combine the strengths of Specialist Models and MLLMs while mitigating their respective weaknesses. We recognize the complementary nature of these architectures: Specialist Models offer efficiency and strong grounding capabilities,

particularly valuable for precise object localization and candidate generation, while MLLMs provide the crucial reasoning abilities needed for disambiguation and complex expression understanding. By strategically integrating these model types, we can achieve superior REC performance with improved computational efficiency compared to relying solely on resource-intensive MLLMs.

Motivated by this insight, we introduce two novel strategies: **Slow-Fast Adaptation (SFA)** and **Candidate Region Selection (CRS)**. SFA routes REC tasks based on their inherent complexity. Simpler tasks, where Specialist Models are adept, are processed efficiently, while more complex tasks requiring reasoning are directed to MLLMs. This dynamic routing optimizes overall computational cost without sacrificing performance. CRS leverages the strong object detection of Specialist Models to generate a set of candidate regions, and then employs an MLLM to reason over these candidates and select the most appropriate one. The subsequent subsections detail each method, elucidating their specific mechanisms and implementation.

A. Slow-Fast Adaptation (SFA)

The Slow-Fast Adaptation (SFA) strategy is designed to optimize both performance and computational efficiency in a training-free manner. As illustrated in Fig. 2(b), SFA operates by utilizing a router to evaluate task complexity and directing tasks to either a fast or slow pathway based on their difficulty level. Given an image and a referring expression, the process begins with target extraction and difficulty level assessment, followed by task routing to the appropriate pathway.

Target Extraction: Given a referring expression, we employ GPT-3.5-turbo to extract the target object by prompting: “Which object does the given expression refer to?”. To minimize hallucinations by GPT, we include three in-context examples in the prompt and require GPT to respond in dictionary format. Detailed prompts are provided in the supplementary material.

Difficulty Level Assessment: We categorize REC tasks into two levels based on the presence of distracting objects in the image:

- **Level 1:** The image contains single instance of the referred object, which can be easily identified based on the category name (e.g., a single white cap in the image, as shown in Fig. 2(b)).
- **Level 2:** The image contains multiple instances of the target category, requiring compositional reasoning to disambiguate the correct target (e.g., identifying a child “positioned to the right of the white cap” among several children, as shown in Fig. 2(b)).

To assess task complexity, we employ a low-cost routing method: object detection is performed on the extracted target category using MM-GDINO-L [66] with a confidence threshold of 0.2. The number of detected objects above this threshold determines the task’s difficulty level. If only one object is detected, the task is classified as Level 1; otherwise, it is classified as Level 2. This approach ensures efficient routing while minimizing computational overhead.

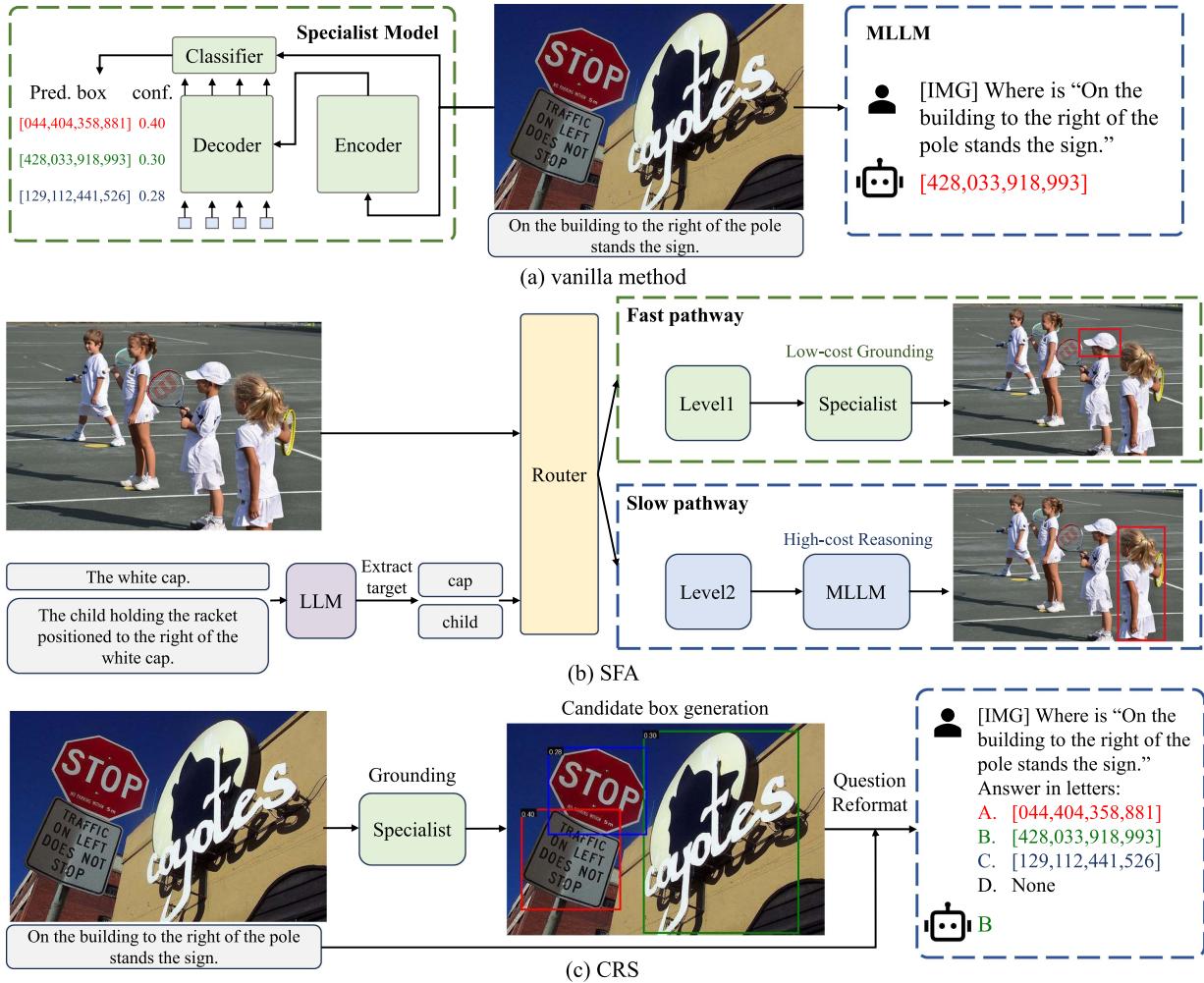


Fig. 2. (a) Vanilla Model: The Specialist Model uses box regression to output multiple candidate boxes with confidence scores and selects the region with the highest confidence. In contrast, MLLMs directly generate object coordinates as text. (b) Slow-Fast Adaptation (SFA): A router categorizes tasks by complexity (Level 1/2). Level 1 tasks are processed using the fast pathway (Specialist Model), while Level 2 tasks are processed using the slow pathway (MLLM). (c) Candidate Region Selection (CRS): The Specialist Model generates multiple candidate bounding boxes, and the MLLM selects the best option in a multi-choice selection task.

Task Routing: Based on the difficulty level, tasks are routed to one of two pathways: Simple tasks are processed using low-cost Specialist Models (e.g., Grounding DINO), which make quick judgments based on the target object’s category. On the other hand, complex tasks are processed using resource-intensive MLLMs, which leverage their superior reasoning capabilities to disambiguate the target object through compositional reasoning.

Notably, the SFA strategy does not require additional training and introduces minimal computational overhead. By dynamically routing tasks based on their complexity, SFA effectively combines the efficiency of Specialist Models with the reasoning ability of MLLMs, achieving higher performance while maintaining computational efficiency.

Focus-enhancement Strategy: In our work, we observe that models often struggle to correctly identify the target object category referred to in an expression, frequently detecting instances from other categories mentioned in the description instead. To mitigate this issue, we propose a focus-enhancement strategy—a simple yet effective method that encourages models to precisely

concentrate on the intended target category derived from the referring expression.

Target-focus Matching for Specialist Models: Fig. 3 illustrates the focus-enhancement strategy for Specialist Models, specifically the target-focus matching mechanism. By default, MM-GDINO [66] computes similarity scores between image tokens and all textual tokens, selecting the proposal with the highest overall score as the output. In contrast, our method refines this process by focusing solely on the similarity scores associated with the target extracted from the referring expression, ensuring that the proposal with the highest target-specific score is selected. This approach proves effective, as cross-modal feature fusion enables the target-specific token to capture both the full semantic context of the expression and the corresponding visual context. By leveraging context-aware, target-specific textual features for matching, our method filters out distracting information while preserving the essential meaning of the expression, thereby improving accuracy. As shown in Fig. 3, by restricting selection to scores associated with the term “bird,” our method

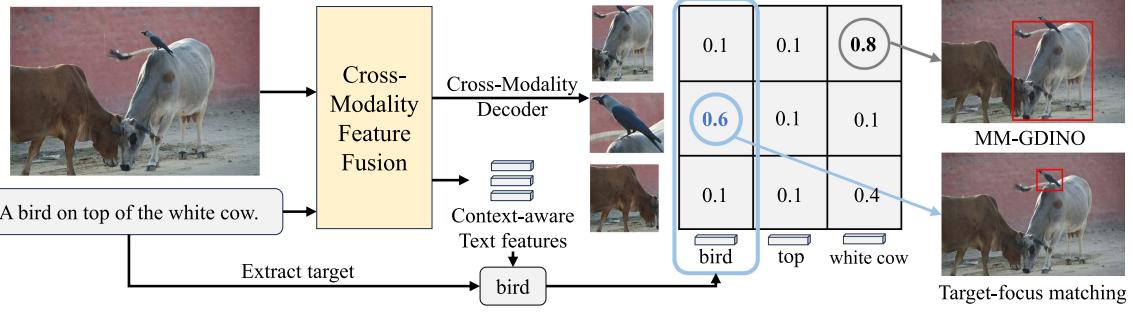


Fig. 3. Illustration of the target-focus matching for Specialist Models. The target is extracted from the referring expression using a LLM. Leveraging the cross-attention mechanism, the target's feature becomes context-aware, enabling more accurate matching with the proposal. This approach helps avoid mismatches with other objects mentioned in the expression, an issue observed in the original solution (e.g., MM-GDINO [66]).

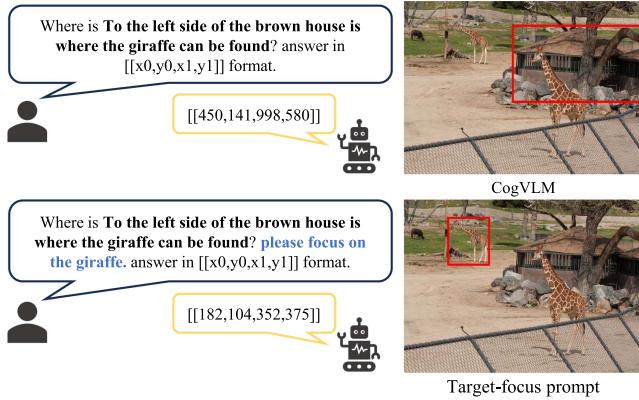


Fig. 4. Target-focus prompt for MLLMs. The default prompt used by CogVLM [55] is “Where is <expr>? answer in $[[x_0, y_0, x_1, y_1]]$ format.” as shown above. To enhance focus, we modify the prompt by appending “please focus on the <target>”, where <target> denotes the target extracted by LLM from the textual expression.

avoids incorrect matches (e.g., where the score for “white cow” is higher) and accurately identifies the intended target.

Target-focus Prompt for MLLMs: Fig. 4 illustrates the focus-enhancement strategy for MLLMs, specifically the target-focus prompt. We add a simple prompt like “please focus on the <target>”, where <target> denotes the extracted target from the expression. As shown in Fig. 4, without the target-focus prompt, CogVLM [55] incorrectly identifies the target as “house”. However, after incorporating the target-focus prompt, the model correctly understands the intended reference “giraffe” and accurately locates its instance on the left side of the image. The purpose of this enhanced prompt is to explicitly emphasize the target of interest within the expression, guiding the model toward more precise interpretation and localization.

B. Candidate Region Selection (CRS)

The Candidate Region Selection (CRS) strategy combines the complementary strengths of Specialist Models and MLLMs in a cost-effective manner. In standard REC tasks, Specialist Models output bounding boxes and typically select the highest-confidence one as the target. However, this approach can fail when the correct target is not the highest-confidence box or

when multiple objects from the same category are present. By exploring the top-k bounding boxes, CRS expands the candidate set, improving the likelihood of selecting the correct target in ambiguous cases. Integrating an MLLM allows for effective disambiguation, enabling the model to reason over the top-k candidates and select the correct one through a multiple-choice format.

CRS Pipeline: As shown in Fig. 2(c), the CRS strategy begins by applying a Specialist Model (e.g., Grounding DINO) for object detection, generating a set of candidate bounding boxes. Non-Maximum Suppression (NMS) with a 0.7 threshold is then applied to remove redundant predictions. From the remaining boxes, the top-k highest-confidence ones are selected as potential targets.

These candidates are passed to the MLLM, which processes them in a multiple-choice format. The MLLM selects the correct bounding box by outputting a single token (e.g., “A”), simplifying the prediction process compared to traditional autoregressive methods that predict bounding box coordinates sequentially. This approach not only reduces inference time but also provides the MLLM with more prior information for accurate target detection.

Instruction Tuning for Multi-Choice Region Selection: While MLLMs are effective for cross-modality reasoning, which is crucial for REC, they may not perform well in the multi-choice region selection task. This is because they were not originally trained for this specific multiple-choice format. To address this, we propose instruction tuning tailored for multi-choice region selection. However, directly tuning the MLLM on the target dataset risks information leakage. Instead, we suggest that the MLLM only needs to acquire the ability to perform region selection, so we construct the instruction tuning dataset based on the external RefCOCO series, a standard REC benchmark that most MLLMs have already encountered during training.

Specifically, we generate a set of top-k bounding boxes for each referring expression using Grounding DINO and retain only those samples where the correct prediction ($\text{IoU} > 0.5$) is within the top-k predictions. These are then shuffled to create candidate choices. This approach allows the MLLM to learn to distinguish between plausible bounding boxes and refine its ability to select the correct one based on both visual and linguistic context.

This instruction tuning strategy equips the MLLM to better handle the multiple-choice format, improving its region selection accuracy while maintaining computational efficiency.

Multi-Choice Instruction Tuning with None Option: A key limitation of MLLMs is their difficulty in rejecting inappropriate queries, largely due to the lack of negative samples in the training data and well-defined refusal mechanisms. To address this, we propose an enhanced instruction tuning approach, sampling 10,000 positive and 2,500 negative samples from the FineCops-ref training set.

Expanding on the standard multiple-choice formulation, we introduce an additional “None” option (e.g., “D. None”) to represent cases where no valid target exists. Our experimental results show that this approach significantly improves the model’s ability to reject negative samples while maintaining high accuracy on positive samples.

V. EXPERIMENTS

In this section, we present the experimental results. In Section V-A, we benchmark FineCops-Ref dataset by evaluating various representative models. The evaluation focuses on the models’ ability to accurately localize the target and reject non-existent objects. In Section V-B, we assess the effectiveness of the proposed Specialist-MLLM collaboration approaches—namely, Slow-Fast Adaptation (SFA) and Candidate Region Selection (CRS)—through comparisons with existing state-of-the-art (SOTA) methods. Finally, in Section V-C, we provide the results of ablation studies to further validate our approach.

A. Benchmarking the FineCops-Ref Dataset

1) Evaluated Models: We evaluate several representative models, including both Specialist Models and MLLMs. The models examined in this study include UNINEXT [42], SimVG [50], MDETR [41], MM-GDINO [43], [66], Shikra [51], Ferret [52], GroundingGPT [53], Lenna [54], CogVLM [55], CogCom [56], InternVL2.5 [65] and Qwen2-vl [64]. We use their open-source checkpoints for evaluation. Additionally, we evaluate GPT4-V [67], which has limited capability in directly outputting bounding boxes. To assess its performance, we combine GPT4-V with the Set-of-Mark (SoM) [58]. Details on the model source and implementation are provided in the Appendix, available online.

2) Evaluation on Positive Data: We evaluate the models using Precision@1 for positive data, as summarized in Table III. The results highlight the importance of categorizing the dataset by difficulty levels, as the performance of most models degrades significantly with increasing complexity. Notably, at difficulty level 3, the majority of models achieve a precision below 50%. An interesting observation is that models like UNINEXT, SimVG, and InternVL2.5, which achieve state-of-the-art performance on the RefCOCO series benchmarks, perform relatively poorly on the FineCops-Ref benchmark. In contrast, Grounding DINO and Qwen2-VL demonstrate strong generalization capabilities, achieving the best results in the Specialist and MLLM categories, respectively.

TABLE III
EVALUATION RESULTS (PRECISION@1) ON POSITIVE DATA. L1, L2, AND L3 REPRESENT LEVEL1, LEVEL2, AND LEVEL3 OF THE GROUND TRUTH LABELS IN FINECOPS-REF, RESPECTIVELY

Model	Positive			
	L1	L2	L3	All
Specialist				
UNINEXT [42]	59.95	43.60	40.98	53.22
SimVG [50]	61.88	47.30	42.04	55.74
MM-GDINO-T [66]	75.11	34.78	35.46	58.87
MDETR [41]	73.25	53.08	46.71	64.80
MM-GDINO-L [66]	85.13	43.54	42.89	68.32
MLLM				
GPT4-V [67] + SoM [58]	55.94	45.94	49.29	52.07
Shikra [51]	64.64	50.29	43.95	58.54
Lenna [54]	73.75	41.92	38.43	60.74
GroundingGPT [53]	71.10	53.64	49.89	63.87
Ferret-13B [52]	70.24	55.99	50.53	64.23
InternVL2.5-8B [65]	70.91	59.67	53.08	66.05
CogVLM [55]	74.59	<u>62.51</u>	57.32	69.46
CogCom [56]	76.27	60.93	<u>59.87</u>	70.03
Qwen2-VL-7B [64]	78.24	66.13	60.72	73.09

The best results are in bold, and the second-best results are underlined.

Specialist performs better on simple REC tasks: At Level 1, the task primarily involves identifying objects based on category names, closely aligning with the requirements of open-vocabulary object detection. Grounding DINO, built on the SWIN-L backbone, achieves an impressive accuracy of 85.13% under this settings. This observation leads to two key insights. First, vision-language models focused on object detection exhibit strong capabilities in basic visual localization and object detection tasks, even in zero-shot scenarios, which is also supported by their superior performance on RefCOCO benchmark which mainly require the model to detect the object without consider the intricate attribute and relation. Second, although MLLMs excel in language understanding and reasoning, their basic object detection abilities still fall short of the standards required for truly general-purpose models.

MLLMs exhibit superior reasoning abilities: At Level 2 and 3, the tasks demand sophisticated language understanding and cross-modal reasoning due to the presence of numerous easily confusable objects in the images. However, most models do not exhibit sufficient capability in this regard. MLLMs based on large language models (LLMs) perform better in this aspect, showcasing their superior compositional reasoning abilities.

3) Evaluation on Negative Data: We evaluate the models using Recall@1 and AUROC for negative data. The AUROC results can be found in Appendix, available online. Specifically, models like MDETR and Lenna that have dedicated object detection modules can generate multiple detection boxes with associated confidence scores, allowing for direct computation of Recall@1 and AUROC. For models that generate coordinates as text using an autoregressive approach, we use the probability of the coordinate tokens to calculate confidence [14], [68]. The

TABLE IV
EVALUATION RESULTS (RECALL@1) ON NEGATIVE EXPRESSIONS

Model	REPLACE						SWAP					
	Object		Attribute		Relation		Object		Attribute		Avg.	
	L1	L2										
Specialist												
MM-GDINO-T [66]	58.84	33.77	50.47	29.96	34.69	31.92	43.89	27.71	43.67	31.97	38.69	
UNINEXT [42]	47.83	33.70	44.66	34.30	39.51	35.61	45.31	37.35	41.69	31.97	39.19	
MDETR [41]	52.89	36.09	50.47	35.92	42.48	40.77	45.89	37.35	44.42	37.70	42.40	
SimVG [50]	43.45	43.20	46.46	44.04	40.33	45.20	41.00	31.93	40.20	50.82	42.66	
MM-GDINO-L [66]	<u>64.23</u>	40.26	55.76	41.52	45.74	43.73	53.02	48.19	49.38	37.70	47.95	
MLLM												
Ferret-13B [52]	38.38	33.01	37.57	34.48	35.58	34.69	38.69	34.94	35.73	35.25	35.83	
GroundingGPT [53]	42.24	35.13	40.14	33.75	37.51	36.72	41.77	39.76	35.24	39.34	38.16	
Shikra [51]	44.99	33.11	41.25	33.03	35.78	39.85	42.27	39.16	39.70	32.79	38.19	
InternVL2.5-8B [65]	38.55	36.16	41.84	37.91	40.92	39.48	39.40	39.16	43.92	38.52	39.58	
CogVLM [55]	53.34	44.02	51.24	48.74	41.22	<u>44.46</u>	47.69	49.40	46.40	40.16	46.67	
CogCom [56]	57.96	44.91	54.65	44.04	45.81	41.70	<u>51.03</u>	43.98	47.39	36.89	46.84	
Lenna [54]	65.88	50.38	<u>58.75</u>	42.96	47.00	43.91	49.94	38.55	<u>49.38</u>	43.44	<u>49.02</u>	
Qwen2-VL-7B [64]	60.63	<u>45.62</u>	59.59	46.93	49.18	42.44	49.13	46.99	50.87	<u>44.26</u>	49.56	

L1 and L2 denote the direct changes related to the target and the changes related to other objects within the expression, respectively.

TABLE V
EVALUATION RESULTS (RECALL@1) ON NEGATIVE IMAGES

Model	REPLACE						SWAP					
	Object		Attribute		Object		Attribute		Flip		Avg.	
	L1	L2										
Specialist												
UNINEXT [42]	48.85	31.62	40.96	30.33	46.91	40.25	37.06	30.66	29.42	37.34		
SimVG [50]	41.69	41.50	43.73	43.15	31.60	38.60	<u>51.75</u>	36.74	39.66	40.94		
MM-GDINO-T [66]	58.46	40.73	44.75	37.61	46.25	51.33	<u>28.67</u>	39.50	40.94	43.14		
MDETR [41]	58.15	42.85	51.70	37.95	48.86	49.49	44.76	44.29	42.22	46.70		
MM-GDINO-L [66]	66.35	49.45	<u>54.93</u>	<u>49.05</u>	<u>55.05</u>	62.63	46.85	<u>45.21</u>	46.48	<u>52.89</u>		
MLLM												
CogCom [56]	32.24	21.55	22.57	20.10	39.74	25.46	18.88	24.13	23.03	25.30		
Shikra [51]	42.57	33.61	36.54	34.26	35.18	38.60	36.36	34.25	37.10	36.50		
GroundingGPT [53]	43.91	36.88	36.31	35.88	39.09	37.17	40.56	37.02	33.05	37.76		
Ferret-13B [52]	41.54	37.46	38.04	36.22	43.00	37.78	39.16	35.27	36.25	38.30		
InternVL2.5-8B [65]	42.96	41.57	39.46	40.80	45.93	44.35	45.45	42.08	40.72	42.59		
Lenna [54]	66.88	<u>51.19</u>	54.38	39.34	47.56	49.08	43.36	33.98	30.92	46.30		
CogVLM [55]	51.11	49.01	43.49	46.10	50.49	<u>53.80</u>	49.65	43.74	37.74	47.24		
Qwen2-VL-7B [64]	69.73	60.65	63.87	58.41	63.52	62.63	54.55	61.88	52.45	60.85		

evaluation results for negative expressions and negative images are shown in Tables IV and V, respectively. The results indicate that existing models generally exhibit low recall@1 performance on negative data, with most values falling below 50%. This suggests that these models have weak rejection capability for non-existent objects. In contrast, the latest model, Qwen2-VL, demonstrates strong generalization ability and achieves the best performance. According to the experimental results, we can draw the following conclusions:

The models struggle with compositional reasoning and exhibit a limited understanding of semantic structures: L1 and L2 represent two levels of modification: L1 refers to direct changes related to the target, including objects, attributes, and relations, while L2 involves changing other object-related elements, which are often more challenging to detect. For the same type of negative data, the recall of L1 is typically significantly higher than that of L2, suggesting that most models can effectively identify simple anomalies, such as changes directly related to

the target. However, for L2 negative data, almost all models perform poorly. This further highlights that the models lack compositional reasoning ability and fail to fully comprehend the semantic structure of sentences.

The models exhibit poor understanding of fine-grained attributes and relations: For negative data at the same level, the recall of changing attributes is generally lower than that of changing objects, while the recall of changing relations is typically significantly lower than the first two. This suggests that the models exhibit relatively stronger object recognition capabilities but weaker understanding of attributes and relations. Specifically, the models perform poorly on negative data involving the “replace relation” and “flip” types, indicating substantial challenges in relation comprehension. Another notable observation is that, compared to directly replacing attributes, the models perform worse when identifying negative data of the “swap attribute” type, highlighting their limitations in accurately associating attributes.

TABLE VI
COMPARISON WITH STATE-OF-THE-ART SPECIALIST AND MLLM METHODS ON THREE REC BENCHMARKS

Method	FineCops-Ref			Ref-Adv			Ref-Reasoning		
	L1	L2	All	L1	L2	All	L1	L2	All
Specialist									
UNINEXT	65.39	39.06	53.22	82.55	69.68	75.59	35.03	16.66	20.80
SimVG	66.86	42.80	55.74	82.90	69.83	75.84	35.79	18.64	22.50
MDETR	74.85	53.10	64.80	81.20	67.38	73.73	46.53	26.62	31.10
MM-GDINO-L	81.20	53.32	68.32	85.37	66.08	74.95	50.94	27.09	32.46
MLLM									
Shikra	66.92	48.79	58.54	79.64	66.62	72.31	40.85	23.83	27.18
GroundingGPT	75.38	50.48	63.87	80.20	66.68	72.89	43.30	21.69	26.55
Ferret-13B	73.77	53.12	64.23	81.61	71.73	76.27	42.27	23.23	27.51
InternVL2.5-8B	76.05	54.40	66.05	80.85	71.73	75.92	44.44	25.97	30.13
CogVLM	73.58	64.68	69.46	79.14	71.98	75.27	45.67	33.10	35.93
CogCom	77.89	60.87	70.03	82.20	71.23	76.27	47.29	28.44	32.68
Qwen2-VL-7B	79.40	65.74	73.09	85.84	74.73	79.83	50.62	31.46	35.77
SFA									
CogVLM(SFA)	81.20	64.68	73.57(+4.11)	85.37	71.98	78.13(+2.86)	50.94	33.10	37.11(+1.18)
CogVLM(SFA†)	83.43	65.69	75.03(+5.57)	85.37	72.98	78.37(+3.10)	52.34	34.37	38.41(+2.48)
Qwen2-VL-7B(SFA)	81.20	65.74	74.06(+0.97)	85.37	74.73	79.62(-0.21)	50.94	31.46	35.84(+0.07)
Qwen2-VL-7B(SFA†)	83.43	66.91	75.61(+2.52)	84.78	75.92	79.97(+0.14)	52.34	31.72	36.36(+0.59)
CRS									
CogVLM(CRS)	79.13	67.20	73.62(+4.16)	81.37	74.58	77.70(+2.43)	51.27	33.97	37.87(+1.94)
Qwen2-VL-7B(CRS)	84.77	70.85	78.33(+5.24)	84.37	77.32	80.56(+0.73)	54.80	36.28	40.45(+4.68)
InternVL2.5-8B(CRS)	86.06	69.88	78.58(+12.53)	87.96	78.42	82.80(+6.88)	55.10	36.05	40.34(+10.21)

L1 and L2 denote Level 1 and Level 2, respectively, as determined by the router Grounding DINO [66] in SFA. † indicates the use of the focus-enhancement strategy in our SFA method.

B. Evaluation of the Proposed Specialist-MLLM Collaboration Approaches

In this section, we compare our proposed Specialist-MLLM collaboration methods with representative REC models, including both Specialist Models and MLLMs. We evaluate the performance using Precision@1. The details for CRS training are provided in Appendix, available online.

1) *Evaluation of Slow-Fast Adaptation (SFA):* As shown in Table VI, Specialist Models generally perform well on Level 1 tasks, where text-to-region feature matching is essential, while MLLM solutions tend to excel in Level 2 tasks, which require more complex reasoning. This observation supports our task-adaptive model selection strategy. Our SFA approach, which combines the strengths of both methods through effective task routing, achieves superior performance compared to using these strategies individually. Incorporating the focus-enhancement strategy further boosts performance, with our SFA† yielding additional improvements. These results demonstrate the effectiveness of our solution in handling REC tasks with varying levels of difficulty, highlighting the broad applicability of our approach.

Importantly, thanks to task-specific model selection, our method achieves high efficiency compared to relying solely on MLLMs. As illustrated in Fig. 5(a), our approach outperforms CogVLM [55] by an absolute 5.6% while using only 62% of the computational cost. When compared to Ferret [52], which maintains similar computational overhead, our method achieves over a 10% performance boost. These results demonstrate that our method strikes an optimal balance between performance and efficiency.

TABLE VII
EFFECTIVENESS OF THE PROPOSED FOCUS-ENHANCEMENT STRATEGY

Model	Method	FineCops-Ref	Ref-Adv	Ref-Reasoning
Specialist				
MDETR [41]	Exp. Foc.	64.80 67.19(+2.39)	73.73 72.12(-1.61)	31.10 34.65(+3.55)
MM-GDINO-L [66]	Exp. Foc.	68.32 71.22(+2.90)	74.95 74.55(-0.40)	32.46 34.28(+1.82)
MLLM				
GroundingGPT [53]	Exp. Foc.	63.87 65.90(+2.03)	72.89 74.33(+1.44)	26.55 28.11(+1.56)
Ferret-13B [52]	Exp. Foc.	64.23 66.10(+1.87)	76.27 78.07(+1.80)	27.51 28.96(+1.45)
CogVLM [55]	Exp. Foc.	69.46 72.11(+2.65)	75.27 77.75(+2.48)	35.93 37.12(+1.19)
InternVL2.5-8B [65]	Exp. Foc.	66.05 69.89(+3.84)	75.92 80.29(+4.37)	30.13 32.29(+2.16)
Qwen2-VL-7B [64]	Exp. Foc.	73.09 75.03(+1.94)	79.83 80.89(+1.06)	35.77 36.16(+0.39)

Exp. denotes the original method, Foc. indicates our focus-enhancement strategy, which can be readily applied to various Specialist and MLLM methods.

Effectiveness of Target-refocus Strategy: To evaluate the generalizability of our proposed focus-enhancement strategy, we apply it to various Specialist and MLLM methods, as shown in Table VII. This strategy helps models focus on the referred target, thereby reducing the misidentification of the target object as another category. The experimental results show that our focus-enhancement strategy consistently improves the performance of both Specialist and MLLM methods across all datasets, demonstrating its effectiveness and broad applicability.

2) *Evaluation of Candidate Region Selection (CRS):* As shown in Table VI, our CRS method consistently achieves improvements across all three benchmarks. Notably, when applied

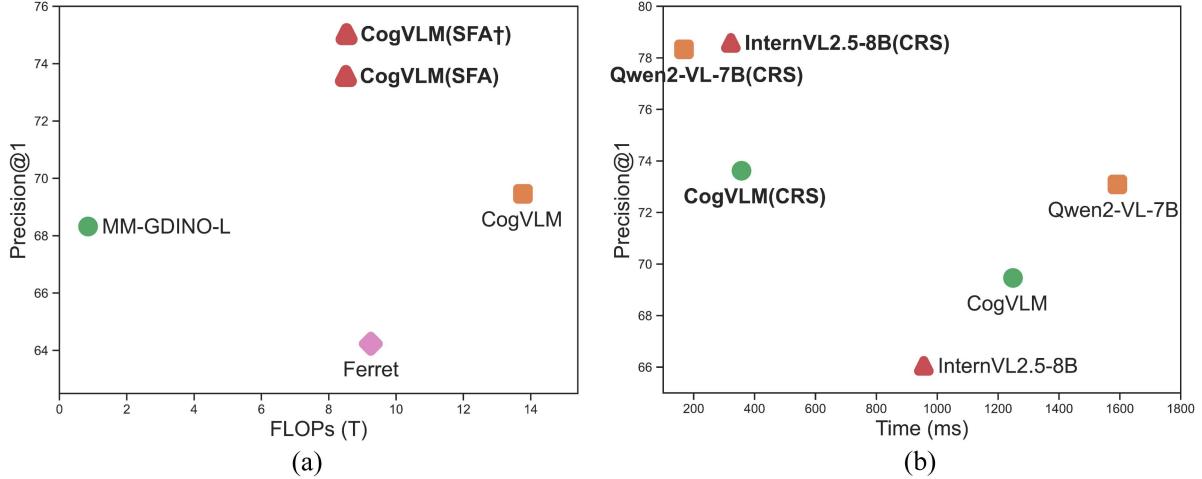


Fig. 5. The effectiveness of the proposed methods. (a) compares the FLOPs (computational cost) of SFA with other methods, including the Specialist Model MM-GDINO-L [66] and MLLMs such as CogVLM [55] and Ferret [52]. (b) compares the inference time of CRS with that of MLLMs themselves.

to InternVL2.5 [65], CRS delivers a remarkable performance boost of nearly 10%. The experimental results demonstrate that advanced models possess substantial cross-modal reasoning capability. Through instruction tuning, we effectively harness these inherent capabilities, enabling the model to achieve stronger grounding performance and fully unleash its potential. These findings underscore both the effectiveness and generalizability of our approach.

It is important to note that, while CRS slightly increases the length of input tokens, it significantly reduces the length of output tokens (from n to 1). In practical applications, this reduction leads to a considerable improvement in inference speed. As illustrated in Fig. 5(b), CRS requires only 10% -30% of the inference time compared to the original method, underscoring the strong practical value of our approach.

As shown in Fig. 6, both Grounding DINO and InternVL2.5 output wrong bounding boxes. However, after applying the CRS method, InternVL2.5, leveraging the candidate boxes provided by Grounding DINO and effectively comprehending the relationship in the expression, successfully selects the correct answer: “B”.

Another advantage of our CRS strategy is its inherent ability to reject non-existent objects by selecting a “None” option. To evaluate this capability, we assess three approaches: (1) using the CRS method without instruction tuning (w/o instruction tuning); (2) instruction tuning an MLLM model with its original REC prompt, followed by the instruction “If no target object exists, please output “None” (Baseline); and (3) instruction tuning the model using the CRS prompt, followed by the instruction “If no suitable option exists, please select the option corresponding to “None” (CRS). We evaluate the models using Accuracy and Recall@1, as summarized in Table VIII. For positive data, Accuracy is equivalent to Precision@1, while for negative data, Accuracy measures the proportion of negative data for which the model correctly selects the “None” option. Recall@1 reflects the average recall score.

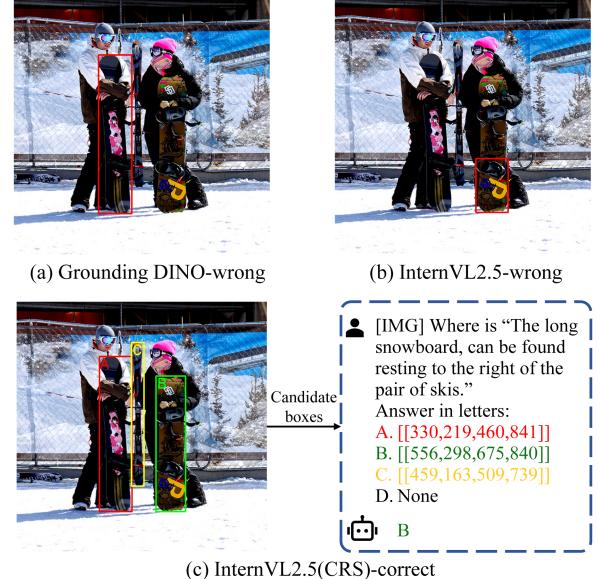


Fig. 6. Visualization of grounding results for Grounding DINO [66], InternVL2.5 [65], and the CRS approach.

TABLE VIII
EVALUATION OF CRS EFFECTIVENESS IN NON-EXISTENT TARGET REJECTION ACROSS THREE SETTINGS: (1) CRS WITHOUT INSTRUCTION TUNING (W/o INSTRUCTION TUNING); (2) INSTRUCTION TUNING WITH THE ORIGINAL PROMPT (BASELINE); AND (3) INSTRUCTION TUNING WITH THE CRS PROMPT (CRS)

Finetune	Model	Accuracy		Recall	
		positive	negative	text	image
w/o instruction tuning	InternVL2.5-8B	73.24	11.80	54.91	54.69
	Qwen2-VL-7B	74.50	5.20	55.37	56.59
	CogVLM	67.83	0.99	45.76	45.84
Baseline	Internvl2.5-8B	69.90	58.32	55.06	56.00
	Qwen2vl-7B	64.45	62.31	55.30	54.97
	CogVLM	70.57	36.52	53.65	55.48
CRS	InternVL2.5-8B	76.74	58.63	69.54	68.31
	Qwen2-VL-7B	79.00	27.08	64.98	68.16

TABLE IX
ABLATION STUDY OF DIFFERENT DETECTION-BASED DIFFICULTY-LEVEL ASSESSMENT METHODS

Method	Accuracy	L2 Recall	FineCops-Ref	Ref-Adv	Ref-Reasoning	Time (ms)
OV-DINO [69]	58.46	34.99	70.39	77.11	35.32	112
ProxyDet [70]	57.82	46.53	70.97	78.21	34.76	79
YOLO-World [71]	67.35	28.70	71.53	77.81	34.15	44
MDETR [41]	61.64	63.34	72.72	79.29	36.18	26
MM-GDINO-L [66]	66.75	66.06	74.06	79.62	37.11	381

Difficulty-level assessment accuracy and L2 recall are evaluated on the FineCops-Ref benchmark, based on ground-truth labels provided.

The experimental results indicate that, without instruction tuning, models frequently fail to select “None,” suggesting that existing models have limited rejection capabilities for non-existent objects. This observation aligns with our previous conclusions. After instruction tuning with the baseline method, models manage to respond with “None,” but their accuracy on positive data drops significantly. This indicates that models fail to properly learn when they should reject. In contrast, instruction tuning with the CRS method improves accuracy on both positive and negative data and demonstrates a substantial advantage in Recall@1. These results validate the effectiveness of the CRS method: the models successfully learn to reject non-existent objects while maintaining precise localization performance for positive data.

C. Ablation Studies

In this section, we present extensive ablation studies on our proposed methods. For SFA, we analyze the impact of different detection-based difficulty-level assessment methods on performance. For CRS, we explore various training configurations, including dataset selection and the number of candidates used for region selection.

1) *Ablation on Difficulty-Level Assessment Methods in SFA:* In SFA, we use an object-detection based strategy to assess task difficulty, which then guides model selection. Specifically, we extract instances that belong to the target category and check if there are multiple instances of that category. To evaluate this approach, we test different object detection methods, selected based on both inference time and detection capacity. The models we choose need to support open vocabulary detection due to the open-world nature of general target extraction.

We compare difficulty-level recognition accuracy and Level 2 recall on FineCops-Ref, which provides ground-truth difficulty-level labels. As shown in Table IX, Grounding DINO [66] achieves superior performance due to its stronger vision backbones. Accuracy indicates the precision of difficulty-level assessment, reflecting the rationality of task complexity division. Level 2 recall measures the proportion of ground-truth Level 2 tasks correctly identified. A higher Level 2 recall suggests that more complex tasks are appropriately assigned to MLLMs. While considering accuracy, it’s crucial to maintain a high Level 2 recall because Specialist Models handle complex tasks poorly, and misassigning these tasks to Specialist Models can significantly degrade SFA performance. Therefore, we select MM-GDINO-L [66] as our detection method, despite

TABLE X
ABLATION STUDY OF VARIATION IN DATASET AND DATASET SIZES IN INSTRUCTION TUNING FOR CRS

Model	Data source	Data size	FineCops-Ref	Ref-Adv	Ref-Reasoning
Qwen2-VL-7B [64]	RefCOCO/+g	-	73.09	79.83	35.77
		2k	74.95	76.40	38.88
		4k	76.49	77.94	39.54
		6k	77.67	79.43	39.93
		8k	78.25	80.37	40.35
		10k	78.33	80.56	40.45
		12k	78.84	81.32	40.68
		FineCops-Ref	10k	81.38	82.64

† denotes the use of multichoice instruction tuning.

TABLE XI
ABLATION STUDY OF VARIATION IN THE NUMBER OF REGION CANDIDATES IN INSTRUCTION TUNING FOR CRS

Model	Top-k	FineCops-Ref	Ref-Adv	Ref-Reasoning
Qwen2-VL-7B [64]	-	73.09	79.83	35.77
	Top-3	77.53	80.16	39.60
	Top-5	78.33	80.56	40.45
	Top-8	77.99	79.45	40.22
	Top-10	78.08	80.08	40.24

† denotes the use of multi-choice instruction tuning.

its higher computational cost, which remains within acceptable limits.

2) *Ablation on Dataset Variation in Instruction Tuning for CRS:* In the CRS approach, we construct datasets from RefCOCO/+g [2], [3], [4] to perform multi-choice instruction tuning. This strategy prevents test data leakage and helps the model adapt to the new output format. Table X shows that performance improves consistently as the dataset size increases from 2 k to 12 k. Additionally, we report results from training on a 10 k dataset constructed from FineCops-Ref. As expected, instruction tuning on FineCops-Ref yields better performance on this dataset. More interestingly, this also leads to improved performance on the other two compositional REC datasets, Ref-Adv and Ref-Reasoning. This highlights the advantage of FineCops-Ref over traditional RefCOCO datasets in terms of enhancing the reasoning capabilities required by MLLMs for handling complex REC tasks.

3) *Ablation on Variation in Number of Region Candidates in Instruction Tuning for CRS:* In the multi-choice instruction tuning process of CRS, the parameter k represents the number of candidate options. Specifically, the bounding boxes generated by Grounding DINO are ranked in descending order according to their confidence scores, and the top- k bounding boxes are selected as the final options. Table XI presents the experimental

results, indicating that the method achieves optimal performance when top-5 candidates are adopted. This is due to the trade-off between recall and the introduction of distractors: while higher top- k values increase recall, they also bring in more irrelevant candidates. Setting $k=5$ strikes a better balance between improving recall and minimizing distractors. Therefore, we adopt a top- k value of 5 for our experimental settings.

VI. CONCLUSION

In this work, we build upon our previously introduced FineCops-Ref dataset, which was designed for fine-grained compositional Referring Expression Comprehension (REC). FineCops-Ref features controllable difficulty levels and carefully constructed negative samples, enabling comprehensive evaluation of REC models in object detection, cross-modal reasoning, and rejection of non-existent targets—capabilities critical for real-world applications. To further advance this task, we proposed a Specialist-MLLM collaboration strategy that leverages the strengths of both object detection and cross-modal reasoning. This strategy is implemented through two approaches: Slow-Fast Adaptation (SFA), which dynamically assigns tasks to the appropriate specialist or MLLM model based on task difficulty, and Candidate Region Selection (CRS), which generates multiple region candidates using specialist models and selects the target through a multi-choice reasoning strategy. Extensive experiments on the FineCops-Ref dataset, along with two other challenging REC datasets, demonstrate the effectiveness of our proposed methods.

REFERENCES

- [1] L. Xiao, X. Yang, X. Lan, Y. Wang, and C. Xu, “Towards visual grounding: A survey,” 2024, *arXiv:2412.20206*.
- [2] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, “Generation and comprehension of unambiguous object descriptions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 11–20.
- [3] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, “Modeling context in referring expressions,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–85.
- [4] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, “Modeling context between objects for referring expression understanding,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 792–807.
- [5] V. Cirik, L.-P. Morency, and T. Berg-Kirkpatrick, “Visual referring expression recognition: What do systems actually learn?,” in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2018, pp. 781–787.
- [6] A. Akula, S. Gella, Y. Al-Onaizan, S.-C. Zhu, and S. Reddy, “Words aren’t enough, their order matters: On the robustness of grounding visual referring expressions,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 6555–6565.
- [7] J. Liu, X. Yang, W. Li, and P. Wang, “FineCops-Ref: A new dataset and task for fine-grained compositional referring expression comprehension,” in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2024, pp. 15440–15457.
- [8] R. Liu, C. Liu, Y. Bai, and A. L. Yuille, “CLEVR-Ref : Diagnosing visual reasoning with referring expressions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4185–4194.
- [9] Z. Chen, P. Wang, L. Ma, K.-Y. K. Wong, and Q. Wu, “Cops-ref: A new dataset and task on compositional referring expression comprehension,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10083–10092.
- [10] S. Yang, G. Li, and Y. Yu, “Graph-structured referring expression reasoning in the wild,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9949–9958.
- [11] D. A. Hudson and C. D. Manning, “GQA: A new dataset for real-world visual reasoning and compositional question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6700–6709.
- [12] Y. Wu, Y. Wei, H. Wang, Y. Liu, S. Yang, and X. He, “Grounded image text matching with mismatched relation reasoning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2964–2975.
- [13] Y. Zeng, Y. Huang, J. Zhang, Z. Jie, Z. Chai, and L. Wang, “Investigating compositional challenges in vision-language models for visual grounding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 14141–14151.
- [14] S. Kurita, N. Katsura, and E. Onami, “RefEgo: Referring expression comprehension dataset from first-person perception of Ego4D,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 15168–15178.
- [15] S. Schulter et al., “OmniLabel: A challenging benchmark for language-based object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 11919–11928.
- [16] R. Shekhar et al., “FOIL it! find one mismatch between image and language caption,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 255–265.
- [17] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, “VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 8253–8280.
- [18] T. Thrush et al., “Winoground: Probing vision and language models for visio-linguistic compositionality,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5228–5238.
- [19] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna, “CREPE: Can vision-language foundation models reason compositionally?,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10910–10921.
- [20] T. Le, V. Lal, and P. Howard, “COCO-counterfactuals: Automatically constructed counterfactual examples for image-text pairs,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 71195–71221.
- [21] A. Kamath, C.-Y. Hsieh, K.-W. Chang, and R. Krishna, “The hard positive truth about vision-language compositionality,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 37–54.
- [22] C.-Y. Hsieh, J. Zhang, Z. Ma, A. Kembhavi, and R. Krishna, “SUG-ARCREPE: Fixing hackable benchmarks for vision-language compositionality,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 31096–31116.
- [23] M. Yarom et al., “What you see is what you read? Improving text-image alignment evaluation,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 1601–1619.
- [24] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, “Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9568–9578.
- [25] J. Zhang, M. Cai, T. Xie, and Y. J. Lee, “CounterCurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2024, pp. 15481–15495.
- [26] Z. Lin, X. Chen, D. Pathak, P. Zhang, and D. Ramanan, “Revisiting the role of language priors in vision-language models,” in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 29914–29934.
- [27] C. Lai, S. Song, S. Yan, and G. Hu, “Improving vision and language concepts understanding with multimodal counterfactual samples,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 174–191.
- [28] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, “Modeling relationships in referential expressions with compositional modular networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4418–4427.
- [29] H. Zhang, Y. Niu, and S.-F. Chang, “Grounding referring expressions in images by variational context,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4158–4166.
- [30] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. Van Den Hengel, “Parallel attention: A unified framework for visual object discovery through dialogs and queries,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4252–4261.
- [31] L. Yu et al., “MAttNet: Modular attention network for referring expression comprehension,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1307–1315.
- [32] X. Liu, Z. Wang, J. Shao, X. Wang, and H. Li, “Improving referring expression grounding with cross-modal attention-guided erasing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1950–1959.
- [33] S. Yang, G. Li, and Y. Yu, “Dynamic graph attention for referring expression comprehension,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4643–4652.
- [34] R. Hong, D. Liu, X. Mo, X. He, and H. Zhang, “Learning to compose and reason with language tree structures for visual grounding,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 684–696, Feb. 2022.

- [35] D. Liu, H. Zhang, F. Wu, and Z.-J. Zha, “Learning to assemble neural module tree networks for visual grounding,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4672–4681.
- [36] P. Wang et al., “Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1960–1968.
- [37] Y. Zhou et al., “A real-time global inference network for one-stage referring expression comprehension,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 1, pp. 134–143, Jan. 2023.
- [38] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, “TransVG: End-to-end visual grounding with transformers,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1749–1759.
- [39] C. Zhu et al., “SeqTR: A simple yet universal network for visual grounding,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 598–615.
- [40] J. Deng et al., “TransVG: End-to-end visual grounding with language conditioned vision transformer,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13636–13652, Nov. 2023.
- [41] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, “MDETR-modulated detection for end-to-end multi-modal understanding,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1760–1770.
- [42] B. Yan et al., “Universal instance perception as object discovery and retrieval,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15325–15336.
- [43] S. Liu et al., “Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 38–55.
- [44] F. Shi, R. Gao, W. Huang, and L. Wang, “Dynamic MDETR: A dynamic multimodal transformer decoder for visual grounding,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1181–1198, Feb. 2024.
- [45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
- [46] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable transformers for end-to-end object detection,” in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–16.
- [47] H. Zhang et al., “DINO: DETR with improved denoising anchor boxes for end-to-end object detection,” in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–19.
- [48] S. Zhao et al., “Generating enhanced negatives for training language-based object detectors,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 13592–13602.
- [49] L. Xiao, X. Yang, F. Peng, Y. Wang, and C. Xu, “OneRef: Unified one-tower expression grounding and segmentation with mask referring modeling,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 139854–139885.
- [50] M. Dai, L. Yang, Y. Xu, Z. Feng, and W. Yang, “SimVG: A simple framework for visual grounding with decoupled multi-modal fusion,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 121670–121698.
- [51] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao, “Shikra: Unleashing multimodal LLM’s referential dialogue magic,” 2023, *arXiv:2306.15195*.
- [52] H. You et al., “Ferret: Refer and ground anything anywhere at any granularity,” in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 1–28.
- [53] Z. Li et al., “GroundingGPT: Language enhanced multi-modal grounding model,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2024, pp. 6657–6678.
- [54] F. Wei, X. Zhang, A. Zhang, B. Zhang, and X. Chu, “Lenna: Language enhanced reasoning detection assistant,” 2023, *arXiv:2312.02433*.
- [55] W. Wang et al., “CogVLM: Visual expert for pretrained language models,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 121475–121499.
- [56] J. Qi et al., “CogCoM: Train large vision-language models diving into details through chain of manipulations,” 2024, *arXiv:2402.04236*.
- [57] Z. Chen et al., “How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites,” *Sci. China Inf. Sci.*, vol. 67, 2024, Art. no. 220101.
- [58] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, “Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V,” 2023, *arXiv:2310.11441*.
- [59] H. Rasheed et al., “GLaMM: Pixel grounding large multimodal model,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 13009–13018.
- [60] H. Zhang et al., “LLaVA-grounding: Grounded visual chat with large multimodal models,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 19–35.
- [61] C. Ma, Y. Jiang, J. Wu, Z. Yuan, and X. Qi, “Groma: Localized visual tokenization for grounding multimodal large language models,” in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 417–435.
- [62] H. Shao et al., “Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 8612–8642.
- [63] J. Li et al., “CoVLM: Composing visual entities and relationships in large language models via communicative decoding,” in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 1–18.
- [64] P. Wang et al., “Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution,” 2024, *arXiv:2409.12191*.
- [65] Z. Chen et al., “Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling,” 2024, *arXiv:2412.05271*.
- [66] X. Zhao et al., “An open and comprehensive pipeline for unified object grounding and detection,” 2024, *arXiv:2401.02361*.
- [67] J. Achiam et al., “GPT-4 technical report,” 2023, *arXiv:2303.08774*.
- [68] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-shot machine-generated text detection using probability curvature,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 24950–24962.
- [69] H. Wang et al., “OV-DINO: Unified open-vocabulary detection with language-aware selective fusion,” 2024, *arXiv:2407.07844*.
- [70] J. Jeong, G. Park, J. Yoo, H. Jung, and H. Kim, “ProxyDet: Synthesizing proxy novel classes via classwise mixup for open-vocabulary object detection,” in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 2462–2470.
- [71] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, “YOLO-world: Real-time open-vocabulary object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16901–16911.

Xuzheng Yang received the BE degree from the University of Electronic Science and Technology of China, in 2024. He is currently working toward the ME degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC). His research interests include multimodal, visual grounding, and computer vision.



Junzhuo Liu received the BE and ME degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2021 and 2024, respectively. He is currently working toward the DEng degree with the School of Computer Science and Engineering, UESTC. His research interests include robustness of deep learning models, multimodal learning.



Peng Wang received the PhD degree from the School of Information Technology and Electrical Engineering, University of Queensland, in 2017. He is currently a professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. He has been actively publishing his research work on top-tier journals and conferences, such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, *CVPR*, and *NeurIPS*. His research interests include computer vision and cross-modal intelligence.

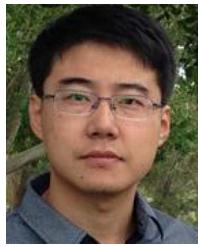




Guoqing Wang (Member, IEEE) received the PhD degree from The University of New South Wales, Australia, in 2021. He is currently with the School of Computer Science and Engineering, University of Electronic of Science and Technology of China. He has authored and co-authored more than 100 scientific articles with top venues, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, ICML, NeurIPS, etc. His research interests include computer vision and remote sensing, with special emphasis on UAV.



Heng Tao Shen (Fellow, IEEE) received the BSc degree with first-class Honours and PhD degree from the Department of Computer Science, the National University of Singapore, in 2000 and 2004, respectively. His research interests mainly include multimedia search, computer vision and artificial intelligence. He is a member of Academia Europaea, fellow of ACM and OSA.



Yang Yang (Senior Member, IEEE) received the PhD degree from The University of Queensland, Brisbane, QLD, Australia, in 2012. He is currently a professor with the University of Electronic Science and Technology of China, Chengdu, China. He has authored and co-authored more than 100 scientific articles at top venues, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, etc. His current research interests include multimedia content analysis, computer vision, and social media analytics.