Connecting the Dots: Training-Free Visual Grounding via Agentic Reasoning

Liqin Luo[1*], **Guangyao Chen**[1*†], **Xiawu Zheng**[4], **Yongxing Dai**[1], **Yixiong Zou**[5], **Yonghong Tian**[1,2,3†]

[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]Peng Cheng Laboratory, China
[3]School of AI for Science, Peking University
[4]Institute of Artificial Intelligence, Xiamen University
[5]School of Computer Science and Technology, Huazhong University of Science and Technology

Abstract

Visual grounding, the task of linking textual queries to specific regions within images, plays a pivotal role in vision-language integration. Existing methods typically rely on extensive task-specific annotations and fine-tuning, limiting their ability to generalize effectively to novel or out-of-distribution scenarios. To address these limitations, we introduce **GroundingAgent**, a novel *agentic visual grounding* framework that operates *without any task-specific fine-tuning*. GroundingAgent employs a structured, iterative reasoning mechanism that integrates pretrained open-vocabulary object detectors, multimodal large language models (MLLMs), and large language models (LLMs) to progressively refine candidate regions through joint semantic and spatial analyses. Remarkably, GroundingAgent achieves an average zero-shot grounding accuracy of 65.1% on widely-used benchmarks (RefCOCO, RefCOCO+, RefCOCOg), entirely without fine-tuning. Furthermore, by substituting MLLM-generated captions with the original query texts, the accuracy at the selection stage alone reaches approximately 90%, closely matching supervised performance and underscoring the critical role of LLM reasoning capabilities. GroundingAgent also offers strong interpretability, transparently illustrating each reasoning step, thus providing clear insights into its decision-making process.

Code — https://github.com/loiqy/GroundingAgent

Introduction

Visual grounding (VG), the task of associating natural language descriptions with specific image regions, is crucial for bridging visual perception and language understanding. This capability underpins numerous downstream tasks, such as visual question answering, human-robot interaction, and interactive image retrieval. Traditional VG benchmarks, including Referring Expression Comprehension (REC) (Yu et al. 2016; Mao et al. 2016) and Referring Expression Segmentation (RES), typically rely on meticulously annotated datasets containing tens of thousands of image-object pairs. This limited scale contrasts sharply with contemporary object detection datasets, which commonly offer millions of annotated instances (Lin et al. 2014; Gupta, Dollar, and Girshick 2019; Zou et al. 2021; Zhang et al. 2024b; Xiao et al. 2025). Consequently, models trained on conventional VG datasets struggle to generalize to open-world scenarios, particularly in zero-shot conditions involving novel or out-of-distribution concepts. Addressing these challenges necessitates sophisticated semantic interpretation, comprehensive scene understanding, and precise spatial localization.

Recently, Transformer-based detectors such as Grounding DINO (Liu et al. 2024) have achieved strong results on visual grounding (VG) benchmarks, notably RefCOCO and RefCOCO+(Yu et al. 2016; Mao et al. 2016). In contrast, multimodal large language models (MLLMs)(Alayrac et al. 2022; Zeng et al. 2025; Chen et al. 2025), while proficient in image captioning and visual question answering due to extensive image-text pre-training (Zhai et al. 2022), exhibit poor localization performance without specialized VG training. For instance, GPT-4o struggles to accurately predict bounding boxes, as illustrated in Figure 1, and language-centric models such as Kosmos-2 (Peng et al. 2023a) significantly underperform compared to detection-based models. Addressing this gap by acquiring additional VG-specific annotations is prohibitively expensive, as precise bounding-box or segmentation annotations are substantially more resource-intensive than image-level captions typically used for MLLM pre-training.

To overcome these challenges, we introduce **GroundingAgent**, a novel *training-free* visual grounding framework empowered by *agentic reasoning*. Unlike conventional methods that rely heavily on costly, task-specific fine-tuning, GroundingAgent capitalizes on the synergistic combination of pre-trained open-vocabulary object detectors, MLLMs, and LLMs. Specifically, GroundingAgent first utilizes an LLM to infer semantically relevant candidate concepts from the given textual query. Subsequently, these concepts guide an open-vocabulary object detector to generate candidate bounding boxes from the input image. Each candidate region is then enriched with detailed visual-semantic descriptions through joint multimodal analysis. These enriched candidates undergo a agentic reasoning process: an LLM progressively refines predictions by considering global image con-

[*]These authors contributed equally.
[†]Corresponding authors.

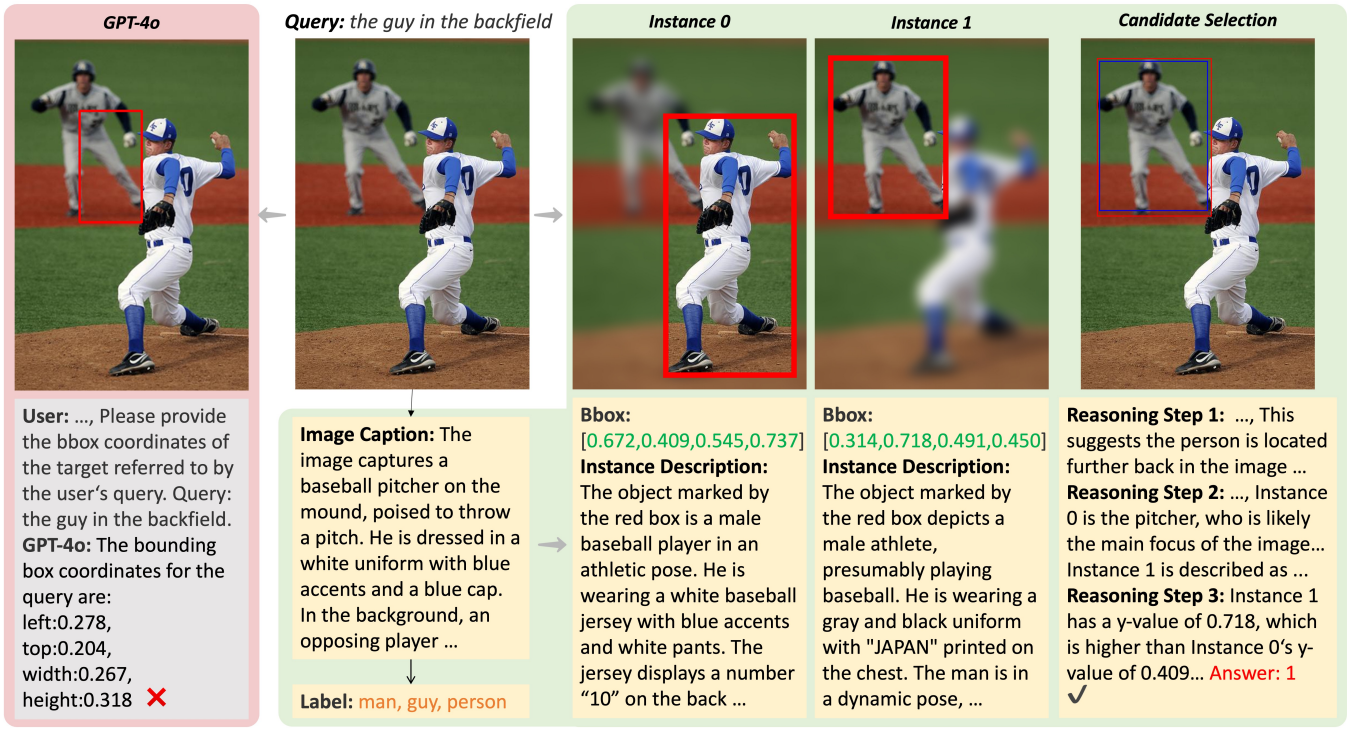

Figure 1: Qualitative comparison and reasoning steps for the visual grounding task. Given the same image and query, the baseline GPT-4o prediction (red box) incorrectly selects the pitcher (left). Our method performs several iterative instance proposals to find the correct object through visual reasoning.

text, candidate semantics, and accumulated reasoning outputs.

We extensively evaluate GroundingAgent on widely-used benchmarks (RefCOCO, RefCOCO+, RefCOCOg), demonstrating superior performance in zero-shot grounding scenarios. GroundingAgent distinguishes itself through several notable advantages: **Firstly**, its open-vocabulary design naturally accommodates novel concepts without being constrained by predefined categories. **Secondly**, by leveraging pretrained MLLMs without any task-specific fine-tuning, GroundingAgent achieves an impressive average accuracy of 65.1% in a fully training-free manner. **Thirdly**, its explicit agentic reasoning pipeline significantly enhances interpretability, transparently revealing each step of the grounding process. **Fourthly**, replacing MLLM-generated captions with query texts during the candidate selection stage boosts selection accuracy to approximately 90%, approaching supervised performance levels. This underscores the critical role of the large language model's reasoning capability in visual grounding.

Our main contributions can be summarized as follows:

- We propose **GroundingAgent**, the first fully training-free visual grounding framework leveraging a structured, agentic reasoning pipeline. It seamlessly integrates pretrained open-vocabulary detectors with multimodal and large language models, entirely avoiding task-specific fine-tuning.
- GroundingAgent achieves state-of-the-art zero-shot performance on standard benchmarks (RefCOCO/+/g), surpassing previous zero-shot methods by significant margins and setting a robust new baseline for training-free visual grounding.
- Our framework demonstrates remarkable interpretability and flexibility, effectively handling complex linguistic instructions involving detailed attributes, spatial relationships, and ambiguous references. Its modular design further allows effortless integration or upgrades of pretrained vision and language models.

Related Work

Visual Grounding. Visual grounding links textual descriptions to image regions and is addressed in both supervised and zero-shot settings. Early methods fine-tuned CNN-based detectors in two-stage (Hong et al. 2022) or one-stage frameworks (Yu et al. 2018; Yang et al. 2020), achieving strong closed-set performance but limited open-set flexibility. The introduction of Vision Transformers (ViTs) (Dosovitskiy et al. 2021) enabled fully transformer-based models like TransVG (Deng et al. 2022), which improve visual–language alignment. More recent work leverages vision–language pre-trained models such as MDETR (Kamath et al. 2021) and Grounding-DINO (Liu et al. 2024) or multi-task architectures like UniTAB (Yang et al. 2022) and OFA (Wang et al. 2022) to boost open-set grounding, at the expense of greater data and compute requirements.

Multimodal Large Language Model. Multimodal large

language models (MLLMs) unify vision and language for a range of tasks. Flamingo (Team 2022) and BLIP-2 (Li et al. 2023) align visual and textual representations via cross-attention or Q-Former modules. These models also support instance-level understanding, enabling visual referring and grounding. For referring, GPT4RoI (Zhang et al. 2024a) and Position-Enhanced Visual Instruction Tuning (Chen et al. 2023a) map text prompts to specific regions. For grounding, Kosmos-2 (Peng et al. 2023b) and Grounding-DINO (Liu et al. 2024) combine box-level detection with grounded pre-training. Explicit approaches inject location tokens (Peng et al. 2023b; Wang et al. 2023a), while implicit methods leverage visual–textual cues (Chen et al. 2023b). Extending these techniques to coherent multi-round dialogues remains a key challenge (Chen et al. 2023b).

Agentic Visual Grounding

Problem Definition. Given an input image I and a natural language query Q, visual grounding aims to locate the target object described by Q by predicting its bounding box $\mathbf{b}_{\text{pred}}$. Formally, let $\mathcal{B}(I)$ denote all possible bounding boxes in I:

$$\mathbf{b}_{\text{pred}} = \arg\max_{\mathbf{b}\in\mathcal{B}(I)} \phi(I, Q, \mathbf{b}),$$

where $\phi(I, Q, \mathbf{b})$ measures the alignment between the visual content in $\mathbf{b}$ and the semantic information in Q. Unlike conventional object detection, where the model simply recognizes and localizes pre-defined object categories, visual grounding requires a joint understanding of both visual cues and the linguistic nuances embedded in Q. Here, $f_{\text{vis}}(I, \mathbf{b})$ and $f_{\text{lang}}(Q)$ denote visual and linguistic representations, respectively, and we define

$$\phi(I, Q, \mathbf{b}) = \text{sim}\big(f_{\text{vis}}(I, \mathbf{b}),\ f_{\text{lang}}(Q)\big),$$

with $\text{sim}(\cdot, \cdot)$ as a similarity metric (e.g., cosine). The challenge is to optimize this objective in a zero-shot setting without task-specific fine-tuning.

GroundingAgent We proposed, GroundingAgent, a training-free framework for zero-shot visual grounding. As shown in Figure 2, a pretrained open-vocabulary detector first suggests candidate bounding boxes. A multimodal large language model (MLLM) then supplies rich semantic descriptions for each region. Finally, a large language model (LLM) reasons step by step over these descriptions, spatial cues, and scene context to pick the box that best matches the textual query. The whole pipeline works without task-specific fine-tuning and provides clear, interpretable reasoning traces.

Candidate Generation For the process for generating candidate target regions, an image caption is firstly generated using an MLLM, which we denote as $C(I)$. This caption provides a comprehensive description of the image content. By concatenating the natural language query Q with the generated caption $C(I)$, we form an enriched textual context that better reflects both the user's intent and the semantic content of the image. Based on this enriched context, a large language model infers a set of semantically relevant candidate target concepts:

$$\mathcal{C}(Q, C(I)) = \{c_1, c_2, \ldots, c_{|\mathcal{C}(Q,C(I))|}\}. \tag{1}$$

For each candidate concept $c \in \mathcal{C}(Q, C(I))$, we employ an open-vocabulary object detector on the input image I to identify corresponding object instances. Specifically, for each concept c, the detector yields a set of candidate bounding boxes:

$$\mathcal{D}_c(I) = \{\mathbf{b}_{c,1}, \mathbf{b}_{c,2}, \ldots, \mathbf{b}_{c,m_c}\}, \tag{2}$$

where $\mathbf{b}_{c,j}$ denotes the j-th bounding box associated with concept c, and m_c is the total number of detections for c. The union of all such detections across candidate concepts forms the overall candidate set:

$$\mathcal{D}(I, Q, C(I)) = \bigcup_{c\in\mathcal{C}(Q,C(I))} \mathcal{D}_c(I). \tag{3}$$

To further refine these candidates, we leverage an MLLM to jointly analyze the global image I and each localized region defined by a bounding box $\mathbf{b}_i$. Let $f_{\text{desc}}(I, \mathbf{b}_i)$ denote the function that generates a detailed description $\mathbf{d}_i$ for the candidate region:

$$\mathbf{d}_i = f_{\text{desc}}(I, \mathbf{b}_i), \tag{4}$$

where $\mathbf{d}_i$ encapsulates both the visual characteristics and semantic attributes of the candidate.

Finally, to prioritize salient objects, the candidate bounding boxes are sorted in descending order of their areas. Denoting the area of a bounding box $\mathbf{b}_i$ as $\text{area}(\mathbf{b}_i)$, the

Algorithm 1 GroundingAgent: Training-Free Visual Grounding via Agentic Reasoning

Require: Image I, natural language query Q
Ensure: Predicted bounding box $\mathbf{b}_{\text{pred}}$

Candidate Generation:
$C(I) \leftarrow \text{MLLM}(I)$ *// Generate global image description*
$\mathbf{T}_{\text{global}} \leftarrow C(I)$ *// Obtain overall semantic context*
$\tilde{Q} \leftarrow \text{Concat}(Q,\ C(I))$ *// Form enriched query context*
$\mathcal{C} \leftarrow \text{LLM}(\tilde{Q})$ *// Infer candidate concepts from enriched query*
for each concept $c \in \mathcal{C}$ **do**
  $\mathcal{D}_c(I) \leftarrow \text{Detector}(I,\ c)$ *// Detect candidate bounding boxes for concept c*
end for
$\mathcal{D}(I, Q, C(I)) \leftarrow \bigcup_{c\in\mathcal{C}} \mathcal{D}_c(I)$
for each candidate bounding box $\mathbf{b}_i \in \mathcal{D}(I, Q, C(I))$ **do**
  $\mathbf{d}_i \leftarrow f_{\text{desc}}(I,\ \mathbf{b}_i)$ *// Generate detailed description*
end for
$\mathcal{D}'(I, Q, C(I)) \leftarrow$
  $\text{Sort}\Big(\text{NMS}(\{(\mathbf{b}_i, \mathbf{d}_i)\}),\ \text{by area}(\mathbf{b}_i)\ \text{(desc)}\Big)$
Candidate Selection:
for each candidate $(\mathbf{b}_i, \mathbf{d}_i) \in \mathcal{D}'(I, Q, C(I))$ **do**
  $r_i \leftarrow \text{LLM}(Q,\ \mathbf{T}_{\text{global}},\ \mathbf{b}_i,\ \mathbf{d}_i)$ *// Evaluate candidate*
end for
$k^* \leftarrow \text{find}\Big(\{r_i\},\ r_i = 1\Big)$
return $\mathbf{b}_{\text{pred}} \leftarrow \mathbf{b}_{k^*}$

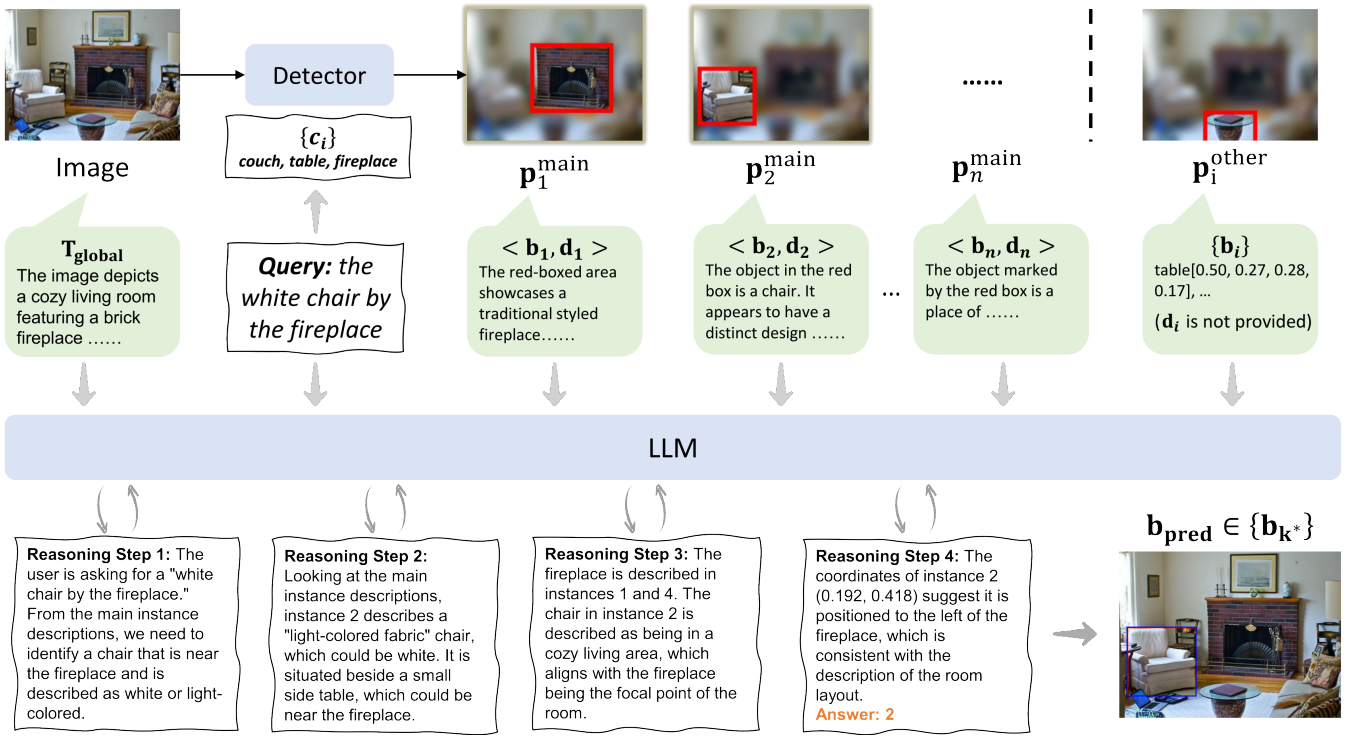

Figure 2: Illustration of our step-by-step reasoning framework for zero-shot referring expression comprehension. Given an input image and a textual query (e.g., *"the white chair by the fireplace"*), the system first extracts a global description $\mathbf{T}_{\text{global}}$ of the scene and generates candidate bounding boxes ($\{\mathbf{b}_i\}$) through an object detector. For each candidate region $\mathbf{b}_i$, an MLLM is employed to generate a fine-grained semantic description $\mathbf{d}_i$, capturing detailed visual attributes and contextual cues. These descriptions, along with the global context and the original query, are passed to an LLM, which performs step-by-step reasoning to refine its understanding of each candidate. In this example, four reasoning steps guide the LLM to identify and confirm the correct bounding box for the white chair, ensuring consistency with the spatial layout and visual attributes described in the query. The final prediction $\mathbf{b}_{\text{pred}}$ is chosen from the candidate set as the best match for the referring expression.

sorted candidate set with non-maximum suppression (NMS) is given by:

$$\mathcal{D}'(I, Q, C(I)) = \text{Sort}\Big(\text{NMS}(\mathcal{D}(I, Q, C(I))), \text{by area}(\mathbf{b}_i)\ (\text{desc})\Big). \tag{5}$$

After this refinement and sorting stage, each candidate is represented as a tuple $(\mathbf{b}_i, \mathbf{d}_i)$, which serves as the foundation for the subsequent grounding process.

Candidate Selection After enriching and sorting the candidate set, we directly select the most appropriate candidate from $\mathcal{D}_{\text{ref}} = \mathcal{D}'(I, Q, C(I))$ based solely on its descriptive caption. For each candidate tuple $(\mathbf{b}_i, \mathbf{d}_i)$, the LLM is provided with the query Q, the global context $\mathbf{T}_{\text{global}}$, the candidate's bounding box $\mathbf{b}_i$, and its detailed description $\mathbf{d}_i$. To further enhance decision-making reliability and interpretability, we incorporate a Chain-of-Thought (CoT) (Wei et al. 2022; Wang et al. 2023b) reasoning module into the process. Specifically, the LLM generates intermediate reasoning steps that explicitly articulate the logical connections between the query, the global context, and the candidate's description. This intermediate chain not only validates the semantic coherence of the candidate with respect to the query but also serves as an internal explanation for the final decision. Without computing any explicit confidence score, the LLM leverages both its semantic understanding and the CoT-derived reasoning to directly judge whether the candidate best corresponds to the query by evaluating its caption:

$$\mathbf{r}_i = \text{LLM}\Big(Q,\ \mathbf{T}_{\text{global}},\ \mathbf{b}_i,\ \mathbf{d}_i\Big) \in \{0, 1\}, \tag{6}$$

where $\mathbf{r}_i = 1$ indicates that the candidate is considered a match. For traditional REC tasks like RefCOCO, which require referring to only one region, the output needs to be constrained to a one-hot format—meaning that there is a unique $\mathbf{r}_i = 1$, and all others are 0. Similarly, this framework can be easily extended to more general referring tasks, where the dataset may include tasks with no referred region or cases where a single referring query corresponds to multiple referred regions(Xia et al. 2024; Liu, Ding, and Jiang 2023). In this way, our caption-based selection strategy not only capitalizes on the LLM's deep semantic understanding but also benefits from the transparent, step-by-step reasoning enabled by the CoT module, thereby obviating the need for iterative refinement or explicit confidence scoring.

The overall process is summarized in Algorithm 1. Our

Method	Zero-shot	RefCOCO			RefCOCO+			RefCOCOg		Avg
		val	testA	testB	val	testA	testB	val	test	
Pseudo-Q (Jiang et al. 2022)	✗	56.0	58.3	54.1	38.9	45.1	32.1	49.8	47.4	47.7
Grounding-Dino (Liu et al. 2024)	✗	50.4	57.2	43.2	51.4	57.6	45.8	67.5	67.1	55.0
MM-G (Zhao et al. 2024)	✗	53.1	59.1	46.8	52.7	58.7	48.4	62.9	62.9	55.6
Kosmos-2 (Peng et al. 2023a)	✗	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.7	52.2
CoVLM (Li et al. 2024)	✗	48.2	53.2	43.2	47.6	50.9	44.2	60.9	61.9	51.3
GLIP (Li* et al. 2022)	✗	50.4	54.3	43.8	49.6	52.8	44.6	66.1	66.9	53.6
REG (Wang et al. 2024)	✗	<u>63.4</u>	68.5	<u>57.6</u>	53.9	60.9	44.9	63.3	63.2	59.5
CPT (Yao et al. 2021)	✓	32.2	36.1	30.3	31.9	35.2	28.8	36.7	36.5	33.5
VGDiffZero (Liu et al. 2023)	✓	28.0	30.3	29.1	28.4	30.8	29.8	33.5	33.2	33.9
ReCLIP (Subramanian et al. 2022)	✓	45.8	46.7	45.2	45.3	48.5	42.7	57.0	56.2	48.4
Red Circle* (Shtedritski et al. 2023)	✓	49.8	58.6	39.9	55.3	63.9	45.4	59.4	58.9	53.9
RelVLA (Han et al. 2023)	✓	52.5	52.7	52.9	50.8	53.4	47.6	61.3	60.9	54.0
FGVP (Yang et al. 2024)	✓	59.6	65.0	52.0	60.0	66.8	49.7	63.3	63.4	60.0
GroundVLP (Shen et al. 2023)	✓	59.1	<u>69.2</u>	48.7	<u>61.8</u>	**70.6**	<u>51.0</u>	**69.1**	**69.0**	<u>62.3</u>
GroundingAgent	✓	**67.1**	**73.3**	**60.1**	**62.4**	<u>67.6</u>	**53.8**	<u>67.9</u>	<u>68.8</u>	**65.1**

Table 1: Comparison with state-of-the-art methods on zero-shot referring expression comprehension (REC) tasks on RefCOCO/+/g dataset. "Zero-shot" here is defined as methods that do not use any task-specific grounding annotations (including manually labeled image-text corresponding regions, grounding labels, etc.) for training or fine-tuning, relying solely on pre-trained model capabilities for inference. The best two results are **bold-faced** and <u>underlined</u>, respectively.

proposed GroundingAgent leverages the complementary strengths of LLMs and open-vocabulary object detectors to perform zero-shot visual grounding. Specifically, GroundingAgent first generates candidate target concepts from the query, then enriches each candidate with detailed semantic and visual descriptions. A two-stage agentic reasoning module subsequently evaluates the spatial and semantic relationships among these candidates, enabling the robust identification of the target object described in Q, all without any additional training.

Experiments

Experiment Setting. To thoroughly evaluate the effectiveness of our proposed agent framework under zero-shot conditions, we conduct experiments on three widely adopted visual grounding benchmarks: RefCOCO/+ (Yu et al. 2016) and RefCOCOg (Mao et al. 2016). These datasets provide a diverse set of referential expressions and corresponding images, facilitating a comprehensive assessment of both bounding-box-level and segmentation-level grounding capabilities. Specifically, we consider the standard task: Referring Expression Comprehension (REC). We measure the top-1 accuracy, where a prediction is deemed correct if the Intersection-over-Union (IoU) between the predicted bounding box and the ground-truth bounding box exceeds 0.5.

Implementation Details. Our training-free visual grounding framework integrates four open-vocabulary detectors—APE (Shen et al. 2024), Grounding DINO (Liu et al. 2024), OWL-ViT (Minderer et al. 2023), and YOLO World (Cheng et al. 2024)—without any task-specific fine-tuning. By default, we use YOLO World, which is not trained on RefCOCO, ensuring an unbiased zero-shot evaluation. Candidate bounding boxes generated by YOLO are sorted by area to prioritize prominent regions. We use Llama-3.2-11B-Vision (Dubey et al. 2024) for generating both global and region-level visual descriptions. Semantic reasoning is performed by DeepSeek-V3 (`0324`) (DeepSeek-AI 2025), which iteratively extracts and refines relevant concepts from queries, global descriptions, and regional semantics. To enhance robustness, candidate regions smaller than 2.5% of the image area are filtered out, and we retain a maximum of 10 primary candidates per image following non-maximum suppression. Additional details, including prompts, normalization strategies, and hyperparameters, are available in the Appendix.

Main Results

We present a comprehensive evaluation of our proposed *GroundingAgent* framework, comparing its performance against state-of-the-art methods for zero-shot referring expression comprehension (REC) tasks on RefCOCO/+/g benchmarks. The detailed comparison results are summarized in Table 1. GroundingAgent consistently achieves superior performance across all benchmark subsets, demonstrating clear advantages over recent zero-shot methods that do not utilize grounding annotations during training. Specifically, our framework attains an average accuracy of **65.1%**, significantly outperforming fully zero-shot competitors such as VGDiffZero (Liu et al. 2023), ReCLIP (Subramanian et al. 2022), and Red Circle (Shtedritski et al. 2023), with accuracy improvements of approximately 12-27% across different subsets. We further emphasize that while REG (Wang et al. 2024) is not a zero-shot method, as it involves training with synthetically generated grounding annotations, introducing implicit supervision and thus deviating from a strictly training-free approach. Despite this implicit training advantage, GroundingAgent outperforms REG on all

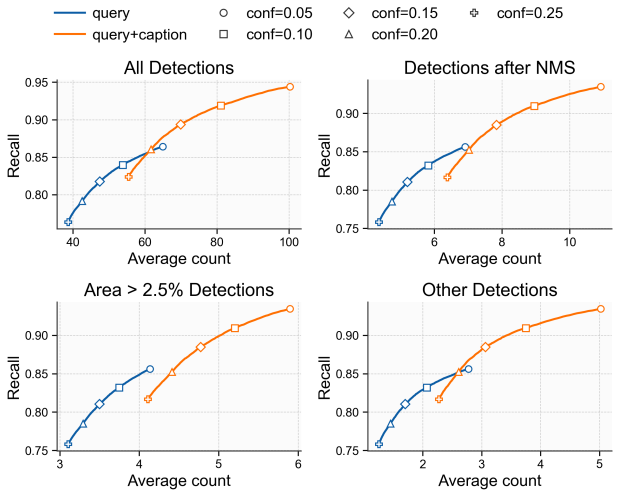

Figure 3: The recall of candidate generation on RefCOCO.

evaluation splits, especially on challenging subsets like RefCOCO+ testB (53.8%) and RefCOCOg test (68.8%), underscoring the genuine zero-shot nature and effectiveness of our fully inference-based method. Moreover, GroundingAgent achieves the best overall performance in terms of average accuracy (65.1%), highlighting its robust generalization capabilities. This consistent performance improvement illustrates the effectiveness of integrating explicit semantic reasoning and structured agentic decision-making, enabling GroundingAgent to effectively interpret complex linguistic queries involving detailed attributes, ambiguous references, and sophisticated spatial relationships. Overall, these results clearly indicate GroundingAgent's effectiveness and its potential as a powerful, flexible baseline for future advancements in the zero-shot visual grounding task.

Candidate Generation

We conduct a detailed analysis to validate the effectiveness of our candidate generation stage, primarily using recall as our evaluation metric. Specifically, a candidate set is considered successful if the ground-truth bounding box is present among the generated candidate boxes.

Robust candidate generation across detectors. First, we investigate the performance differences when employing various open-vocabulary detection models. The results, presented in Table 2, indicate that almost all detection models achieve recall rates around 95%, underscoring the robustness of our candidate generation process. It is important to highlight that these results are obtained solely from bounding box outputs generated by detection models. We do not leverage any additional scoring, confidence metrics, or supplementary information from these detection models. Additionally, the inputs provided to these detection models are limited strictly to candidate vocabularies extracted via our proposed method, demonstrating the strength of the vocabulary generation step.

Global captions are essential for precise vocabularies. To further examine the influence of different textual inputs during candidate vocabulary generation, we conduct experiments using two distinct input scenarios: using only the textual query versus combining both the textual query and a global image caption. The comparative results are illustrated in Figure 3. It is evident from the figure that removing caption information leads to a significant drop in recall, highlighting the importance of global contextual captions. This suggests that incorporating global image descriptions effectively constrains the vocabulary generation process within the LLM, mitigating semantic divergence. Consequently, the captions guide the LLM toward generating more precise and relevant candidate vocabularies, directly enhancing the downstream grounding performance.

Candidate Selection

Caption quality is the main bottleneck. Table 2 demonstrates a noticeable decline in overall performance following the candidate selection step. To thoroughly investigate this phenomenon, we conducted an ablation experiment where the caption descriptions generated by the MLLM for target instances were directly replaced with their original textual queries. Table 2 also provides a detailed comparison illustrating that when using either the original query or a closely related textual description (Query+), the LLM achieves approximately 90% accuracy. This result suggests that the candidate selection performance drop does not originate from our LLM-based selection process. Instead, it is largely due to inaccuracies in the caption generated by the MLLM.

Performance Alignment with Supervised SOTA. The performance of our framework after substituting MLLM-generated captions with Query+ (85.0% average accuracy) closely approaches the accuracy of SOTA fine-tuned model (Zheng et al. 2025) (84.1% average accuracy). This alignment suggests that when provided with query-enhanced contextual cues, our training-free framework can achieve results comparable to methods specifically optimized for visual grounding through task-specific fine-tuning, implicitly validating its ability to leverage semantic alignment without explicit supervision. Furthermore, replacing captions with the original query directly yields an average accuracy of 90.6%, which matches the performance of SOTA pretrained model (Team 2025b) (90.3% average accuracy) that are pretrained on massive vision-language datasets. This equivalence demonstrates that our framework's core reasoning mechanism—when unimpaired by MLLM caption noise—can fully harness the semantic understanding capabilities of pre-trained models, confirming the effectiveness of its modular design in bridging vision and language without task-specific training.

Reasoning ability matters more than model size. We further explored the impact of different LLMs on selection performance through additional ablation studies (Table 3). The results reveal significant performance disparities across LLMs: advanced models like DeepSeek-R1 achieve 75.9% and 60.3% accuracy on RefCOCO testA and testB, respectively, outperforming DeepSeek-V3. Notably,

Stage	Method	RefCOCO			RefCOCO+			RefCOCOg		Avg
		val	testA	testB	val	testA	testB	val	test	
Candidate Generation	APE (Shen et al. 2024)	98.6	98.7	97.9	98.1	98.5	98.0	98.4	98.6	98.3
	GroundingDINO (Liu et al. 2024)	98.3	98.7	97.6	97.9	99.0	97.8	98.2	98.3	98.2
	OWL (Minderer et al. 2023)	95.7	96.3	92.6	95.7	95.7	93.0	95.0	95.3	94.9
	YOLO-World (Cheng et al. 2024)	94.4	96.7	91.1	94.2	95.7	91.3	93.2	93.8	93.8
Candidate Selection	Caption	67.1	73.3	60.1	62.4	67.6	53.8	67.9	68.8	65.1
	Caption → Query+	82.9	83.7	78.8	85.9	82.5	86.4	89.6	90.2	85.0
	Caption → Query	91.4	91.4	86.6	91.4	89.9	90.2	91.8	92.4	90.6
Supervised SOTA	Fine-tuned Model (Zheng et al. 2025)	90.5	91.7	88.0	80.1	84.6	72.6	82.5	82.9	84.1
	Pre-trained Model (Team 2025b)	92.7	94.6	89.7	88.9	92.2	83.7	89.9	90.3	90.3

Table 2: Ablation study on candidate generation and selection strategies. We report accuracy (%) for each open-vocabulary detector independently, as well as the results after applying the candidate selection stage.

LLM	RefCOCO	
	testA	testB
DeepSeek-V3 (DeepSeek-AI 2025)	73.3	60.1
DeepSeek-R1 (DeepSeek-AI 2025)	75.9	60.3
Llama3.1-8B (Dubey et al. 2024)	55.0	44.0
DeepSeek-R1-Llama-8B	59.7	47.7
Qwen2.5-7B (Team 2025a)	52.0	41.6

Table 3: Ablation study on LLM in candidate selection stage.

even among models with comparable parameter scales, performance varies substantially based on reasoning capabilities. For example, DeepSeek-R1-Llama-8B—which incorporates explicit reasoning training—outperforms the base Llama3.1-8B by 4.7-3.7 percentage points. This indicates that enhancing reasoning abilities through targeted training can effectively boost performance even without increasing model size, highlighting the critical role of structured reasoning in the selection process. In contrast, LLMs with weaker understanding and reasoning capabilities, such as Qwen2.5-7B, exhibit substantially lower performance.

Robust LLMs fix issues. For the reasoning steps, we conducted a qualitative analysis on the RefCOCO dataset. For LLMs with weak long-text processing, poor instruction-following, and limited reasoning performance, two issues may arise: first, **not reasoning**, meaning the LLM directly provides an answer without the necessary reasoning steps; second, **ambiguity**, meaning the descriptions generated by the LLM are too vague or unclear, sometimes even exhibiting inherent hallucinations in long text processing, which makes subsequent selection steps difficult. For high-performing LLMs, the model provides reasonable reasoning steps and clear descriptive information, effectively guiding the subsequent selection process. The average number of reasoning steps is 3.4, indicating that the LLM considers multiple steps during reasoning. We also found that, during the reasoning process, the LLM integrates the global semantic description of the image to infer or correct unreasonable text queries, thereby enhancing interpretability and improving robustness.

Further Analysis

Failure Analysis. Our method demonstrates robust performance, with quantitative analysis revealing low rejection rates on the RefCOCO+ benchmark (0.77% on val, 0.73% on testA, and 1.69% on testB), confirming the effectiveness of our rejection-aware selection strategy. Primary error sources include inaccurate descriptions from multimodal generation artifacts and incorrect rejections with ambiguous referring expressions. Detailed failure case analysis is provided in the appendix.

Better explainability. To demonstrate interpretability, we present examples of accepted and rejected bounding boxes in Figure 1. First, our visual prompt guides attention by outlining each candidate region in red and blurring the background. Then, our LLM-based selection strategy provides clear reasoning for each decision, explaining why candidates are accepted or rejected. When candidates share visual or spatial similarities, the framework evaluates their descriptions and positions to justify the final choice. This clarity enhances transparency, simplifies debugging, and reveals the model's decision process.

Conclusion

We presented *GroundingAgent*, a training-free visual grounding framework that integrates pretrained models without task-specific fine-tuning. It achieves competitive zero-shot results on standard benchmarks and effectively interprets complex linguistic queries involving attributes, spatial relations, and ambiguity. The explicit reasoning pipeline enhances interpretability, while the modular design enables easy model substitution and future scalability. Our analysis shows that remaining limitations mainly arise from the fine-grained visual abilities of current MLLMs, suggesting clear directions for progress as multimodal modeling improves. Similar training-free, reasoning-based frameworks may also generalize well to other domains that require transparent and reliable decision-making.

Acknowledgements

The authors gratefully acknowledge support from the Shenzhen High-Level Talent Team Program and the Shenzhen Science and Technology Innovation Commission (Grant No. KQTD 20240729102051063); the National Natural Science Foundation of China (Grant Nos. 62332002, 62027804, 61825101, and 62402015); the China Postdoctoral Science Foundation (Grant No. 2024M750100); and its Postdoctoral Fellowship Program (Grant No. GZB20230024). Computational resources were provided by Pengcheng Cloudbrain.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Chen, C.; Qin, R.; Luo, F.; Mi, X.; Li, P.; Sun, M.; and Liu, Y. 2023a. Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models. arXiv:2308.13437 [cs].

Chen, G.; Horstmann, K.; Wang, Z.; and You, F. 2025. Automated Essential Concept Discovery for Few-Shot Out-of-Distribution Detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3964–3974.

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

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*.

DeepSeek-AI. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2022. TransVG: End-to-End Visual Grounding with Transformers. arXiv:2104.08541 [cs].

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs].

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv. org*.

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

Han, Z.; Zhu, F.; Lao, Q.; and Jiang, H. 2023. Zero-shot Referring Expression Comprehension via Structural Similarity Between Images and Captions. *arXiv preprint arXiv:2311.17048*.

Hong, R.; Liu, D.; Mo, X.; He, X.; and Zhang, H. 2022. Learning to Compose and Reason with Language Tree Structures for Visual Grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2): 684–696.

Jiang, H.; Lin, Y.; Han, D.; Song, S.; and Huang, G. 2022. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15513–15523.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR – Modulated Detection for End-to-End Multi-Modal Understanding. arXiv:2104.12763 [cs].

Li, J.; Chen, D.; Hong, Y.; Chen, Z.; Chen, P.; Shen, Y.; and Gan, C. 2024. CoVLM: Composing Visual Entities and Relationships in Large Language Models Via Communicative Decoding. In *The Twelfth International Conference on Learning Representations*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs].

Li*, L. H.; Zhang*, P.; Zhang*, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; Chang, K.-W.; and Gao, J. 2022. Grounded Language-Image Pre-training. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

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

Liu, X.; Huang, S.; Kang, Y.; Chen, H.; and Wang, D. 2023. Vgdiffzero: Text-to-image diffusion models can be zero-shot visual grounders. *arXiv preprint arXiv:2309.01141*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*.

Minderer, M.; Gritsenko, A.; Houlsby, N.; and Minderer, M. 2023. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36: 72983–73007.

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

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

Shen, H.; Zhao, T.; Zhu, M.; and Yin, J. 2023. GroundVLP: Harnessing Zero-shot Visual Grounding from Vision-Language Pre-training and Open-Vocabulary Object Detection. arXiv:2312.15043.

Shen, Y.; Fu, C.; Chen, P.; Zhang, M.; Li, K.; Sun, X.; Wu, Y.; Lin, S.; and Ji, R. 2024. Aligning and prompting everything all at once for universal visual perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13193–13203.

Shtedritski, A.; Rupprecht, C.; Vedaldi, A.; and Shtedritski, A. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11987–11997.

Subramanian, S.; Merrill, W.; Darrell, T.; Gardner, M.; Singh, S.; and Rohrbach, A. 2022. ReCLIP: A Strong Zero-Shot Baseline for Referring Expression Comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5198–5215.

Team, F. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv:2204.14198 [cs].

Team, Q. 2025a. Qwen2.5 Technical Report. arXiv:2412.15115.

Team, Q. 2025b. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. arXiv:2202.03052 [cs].

Wang, S.; Kim, D.; Taalimi, A.; Sun, C.; and Kuo, W. 2024. Learning visual grounding from generative vision and language model. *arXiv preprint arXiv:2407.14563*.

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; and Dai, J. 2023a. VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. arXiv:2305.11175 [cs].

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023b. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171.

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.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. GSVA: Generalized Segmentation via Multimodal Large Language Models. arXiv:2312.10103 [cs].

Xiao, D.; Chen, G.; Peng, P.; Huang, Y.; Zhao, Y.; Dai, Y.; and Tian, Y. 2025. When Every Millisecond Counts: Real-Time Anomaly Detection via the Multimodal Asynchronous Hybrid Network. *arXiv preprint arXiv:2506.17457*.

Yang, L.; Wang, Y.; Li, X.; Wang, X.; and Yang, J. 2024. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 387–404. Springer.

Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Ahmed, F.; Liu, Z.; Lu, Y.; and Wang, L. 2022. UniTAB: Unifying Text and Box Outputs for Grounded Vision-Language Modeling. arXiv:2111.12085 [cs].

Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*.

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. arXiv:1801.08186 [cs].

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, 69–85. Springer.

Zeng, Y.; Wu, H.; Nie, W.; Chen, G.; Zheng, X.; Shen, Y.; Peng, J.; Tian, Y.; and Ji, R. 2025. From Objects to Events: Unlocking Complex Visual Understanding in Object Detectors via LLM-guided Symbolic Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 24380–24391.

Zhai, X.; Kolesnikov, A.; Houlsby, N.; and Beyer, L. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12104–12113.

Zhang, S.; Sun, P.; Chen, S.; Xiao, M.; Shao, W.; Zhang, W.; Liu, Y.; Chen, K.; and Luo, P. 2024a. GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. arXiv:2307.03601 [cs].

Zhang, Z.; Chen, G.; Zou, Y.; Huang, Z.; Li, Y.; and Li, R. 2024b. Micm: Rethinking unsupervised pretraining for enhanced few-shot learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7686–7695.

Zhao, X.; Chen, Y.; Xu, S.; Li, X.; Wang, X.; Li, Y.; and Huang, H. 2024. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*.

Zheng, S.; Zhao, P.; Zheng, Z.; He, P.; Cheng, H.; Cai, Y.; and Huang, Q. 2025. Look Around Before Locating: Considering Content and Structure Information for Visual Grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2): 1656–1664.

Zou, Y.; Zhang, S.; Chen, G.; Tian, Y.; Keutzer, K.; and Moura, J. M. 2021. Annotation-efficient untrimmed video action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, 487–495.