REF-ADV: EXPLORING MLLM VISUAL REASONING IN REFERRING EXPRESSION TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Referring Expression Comprehension (REC) links language to region level visual perception. Standard benchmarks (RefCOCO, RefCOCO+, RefCOCOg) have progressed rapidly with multimodal LLMs but remain weak tests of visual reasoning and grounding: (i) many expressions are very short, leaving little reasoning demand; (ii) images often contain few distractors, making the target easy to find; and (iii) redundant descriptors enable shortcut solutions that bypass genuine text understanding and visual reasoning. We introduce Ref-Adv, a modern REC benchmark that suppresses shortcuts by pairing linguistically nontrivial expressions with only the information necessary to uniquely identify the target. The dataset contains 5k expressions on real images (1k human authored, 4k human verified), curated with hard distractors and annotated with reasoning facets including negation. We conduct comprehensive ablations (word order perturbations and descriptor deletion sufficiency) to show that solving Ref-Adv requires reasoning beyond simple cues, and we evaluate a broad suite of contemporary multimodal LLMs on Ref-Adv. Despite strong results on RefCOCO, RefCOCO+, and RefCOCOg, models drop markedly on Ref-Adv, revealing reliance on shortcuts and gaps in visual reasoning and grounding. We provide an in depth failure analysis and aim for Ref-Adv to guide future work on visual reasoning and grounding in MLLMs.

1 INTRODUCTION

Referring expression comprehension (REC) is the task of grounding a natural language expression to a specific region in an image (Mao et al., 2016; Kazemzadeh et al., 2014; Yu et al., 2016). It has important applications in real world systems and downstream tasks, and it has become a key benchmark for evaluating multimodal large language models (MLLMs) because it probes fine grained correspondence between language and vision. Recent MLLMs (Google, 2025; Cloud, 2025; Laboratory, 2025), both closed source and open source, have made substantial progress, achieving over 90% accuracy on classic REC benchmarks, i.e., RefCOCO(+/g) (Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016).

Despite this near saturated performance, we identify critical limitations of the classic REC benchmarks that motivate a modern benchmark capable of more challenging and comprehensive evaluation of MLLMs. We view modern REC for MLLMs as a multistep reasoning task with two coupled components: (1) textual reasoning—understanding the referring expression, identifying the target, and identifying its descriptors; and (2) visual reasoning—searching for candidates and establishing correspondence between descriptors and image regions. The order of these steps can vary across models, but a meaningful benchmark should require both textual and visual reasoning. From this perspective, we highlight the following limitations of RefCOCO(+/g).

First, ***most of the referring expressions are extremely short***, as shown in Figure 1. For RefCOCO and RefCOCO+, the average expression length is around 3 words. Such short expressions lead to two issues: (1) minimal linguistic effort is required, and (2) they typically entail less visual reasoning because fewer descriptors must be verified in the image. Second, ***there are few distractors in the images in RefCOCO(+/g)***, as shown in Figure 2 (b), with most cases of only 1 distractors. Here we define a distractor as an object of the same category as the target but a different instance. When few distractors exist, the task requires far less textual and visual reasoning: models need only infer the

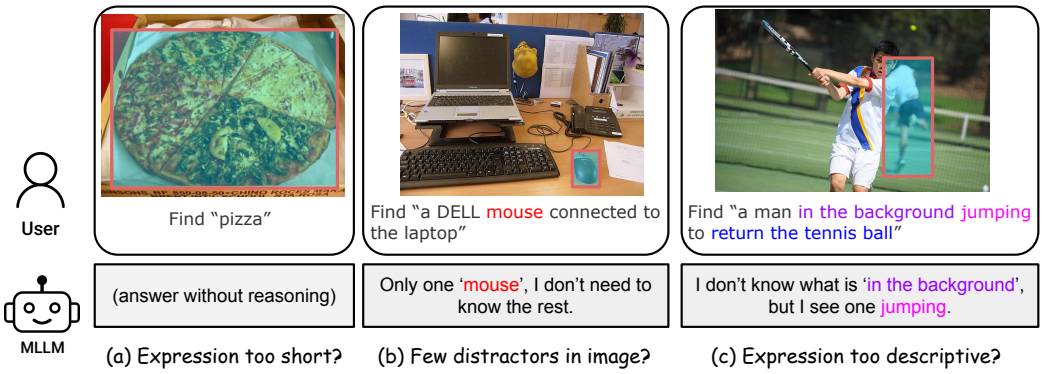


Figure 1: Common limitations of classic referring expression benchmarks that reduce the reasoning challenge. These include very short expressions, few visual distractors, and overspecified descriptors that enable shortcut matching without requiring genuine reasoning. The cyan box highlights the ground truth region.

target category and select from a small set of candidates. Figure 2 (b) reveals a negative correlation between the number of distractors and model performance.

It is worth noting that for reasoning assessment, ***task difficulty does not monotonically increase with referring expression length due to "grounding shortcuts"***. These shortcuts occur when a long, descriptive expression is paired with few distractors, rendering many descriptors redundant. Consequently, a model can localize the target by matching only a subset of descriptors, which can paradoxically lead to higher accuracy for longer expressions, as illustrated in Figure 2 (a). This highlights the need for modern REC benchmarks to mitigate such shortcuts by designing expressions that are concise and carefully balanced against the available distractors.

Meanwhile, prior work has acknowledged aspects of these limitations: Wei et al. (2024); Chen et al. (2024) point out the length limitations of RefCOCO(+/g), and Chen et al. (2020) highlights the lack of distractors. However, the proposed datasets also raise new concerns. The former introduces REC data with average length ≥ 90 words, which may be unnatural and, more importantly, enable numerous shortcuts since the numbers of descriptors and distractors are heavily imbalanced. The latter proposes settings including referring from a set of images, which shifts away from the classic REC setting, and the referring expressions are sampled from GQA (Hudson & Manning, 2019) scene graphs with fixed templates, reducing naturalness.

We therefore aim to build a REC benchmark that preserves the classic REC setting and natural expressions while substantially increasing the reasoning challenge aligned with the capabilities of modern LLMs. To this end, we introduce Ref-Adv, a modern REC benchmark that avoids short reasoning paths and imposes both reasoning and grounding challenges on contemporary MLLMs. To validate and ensure the quality of the benchmark, we conduct comprehensive in depth ablation studies in section 2 to explore what makes a rigorous modern REC benchmark and compare its reasoning and grounding difficulty with RefCOCO(+/g). Lastly, in section 3, we evaluate 13 contemporary MLLMs on Ref-Adv, both closed source and open source. We report changes in performance with and without Chain-of-Thought (CoT) (Wei et al., 2022) and provide in depth analyses. We believe these results demonstrate the value of Ref-Adv, offer new insights into the capabilities of current MLLMs, and can help guide future research on visual reasoning and REC tasks.

2 The Ref-Adv Dataset

2.1 Data Source

We sample from the validation and test splits of COCO (Lin et al., 2014) and OpenImages v7 (Kuznetsova et al., 2020). We filter the images and only use those with panoptic instance annotations, since this is important for our later pipeline. For the bounding box annotations, we convert

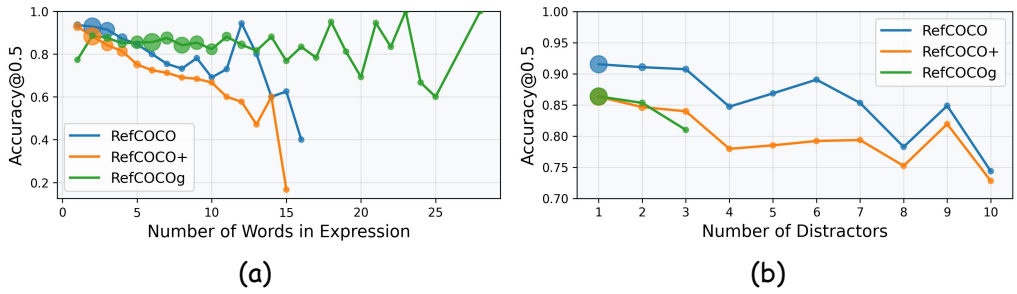


Figure 2: Accuracy@0.5 (IoU $\geq$ 0.5) of Qwen on the RefCOCO/+/g validation sets. Marker size is proportional to the number of samples in each bin. (a) is the Acc@0.5 on number of words in expressions, (b) is on distractor count. We can see most cases have short expressions and few distractors.

Table 1: Basic statistics of the validation+test sets of RefCOCO, RefCOCO+, RefCOCOg, and Ref-Adv (Ours). The instance size is represented by its square root. Avg. length: average length of annotations. Vocab.: vocabulary size. Avg. distractors: average number of same category distractors per image. Negation ratio: percentage of expressions using explicit negation.

Benchmark	Images	Instances	Avg. Length	Avg. Distractors	Negation Ratio	Instance Size	Vocab.
RefCOCO 2014	3,000	7,596	3.6	3.99	0.99%	105–607	3,525
RefCOCO+ 2016	3,000	7,578	3.6	3.96	3.36%	105–607	4,387
RefCOCOg 2016	3,900	7,596	8.4	1.64	1.41%	83–610	5,050
Ref-Adv (Ours)	2,833	5,000	11.5	4.01	21.25%	30-607	5,308

all to using the absolute coordinates in the format of [x1, y1, x2, y2]. The input for our data pipeline is the image, the bounding box annotations and category name of each instance, and we will output the referring expression paired with the target instance.

2.2 Collection Guidelines

As shown in Figure 1, we aim to collect referring data that requires visual reasoning, avoids shortcut solutions, and challenges models. Based on these observations, we propose the following guidelines to mitigate these limitations and yield cases requiring advanced reasoning.

Distractor Pressure Distractors are instances of the same category as the target but different instances. To avoid easy grounding based solely on the target category, we select images that have at least 3 candidate instances of the same category as the target, based on the instance annotations of each dataset.

Language Complexity RefCOCO(+/g) has an average expression length of around 3 words, which limits language complexity and requires much less visual reasoning. Meanwhile, fixed templates that extract referring information from scene graphs limit diversity in the referring expressions. Therefore, we employ LLMs (e.g., GPT-4o) with carefully designed pipelines to generate more natural and diverse referring expressions while maintaining linguistic complexity.

Hard Distractors Simply increasing the number of distractors and the length of the referring expression does not necessarily make the task more challenging because of the "grounding shortcut" illustrated in Figure 1 (c). To reduce such shortcuts (i.e., reliance on redundant descriptors), we ensure the presence of "hard distractors" in the images, defined as distractors that partially match, but do not exactly satisfy, the referring expression. Identifying such pairs and composing expressions around them is central to our data collection process.

Manual Check It is laborious and time-consuming to manually select images with hard distractors and generate the referring expressions, so we use LLMs to assist generation. However, LLMs can

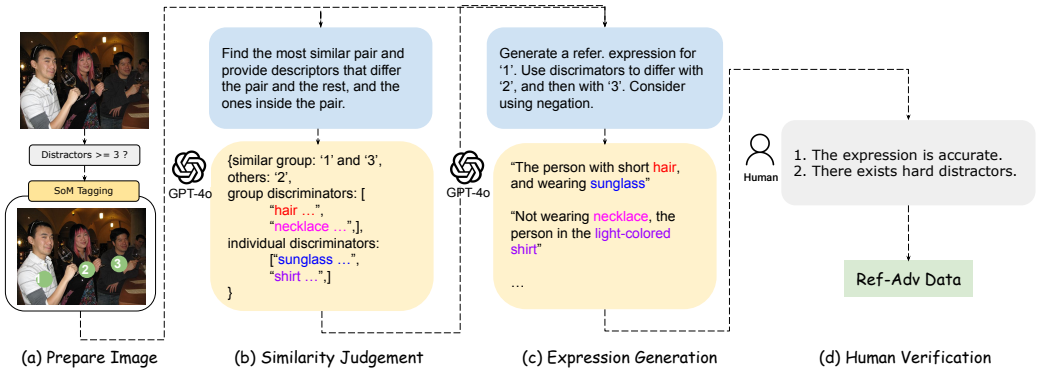


Figure 3: LLM-authored data curation pipeline for Ref-Adv. (a) Prepare Image: filter images, ensure ≥ 3 distractors, and add number tags to candidate instances. (b) Similarity Judgement: use GPT-4o to identify the most similar pair and elicit group-level and instance-level discriminators. (c) Expression Generation: compose minimally sufficient referring expressions using discriminators and optional negation. (d) Human Verification: verify expression accuracy and confirm the existence of hard distractors before inclusion.

make mistakes or hallucinate. To ensure accuracy, we perform a human verification pass to confirm the existence of hard distractors and the correctness and unambiguity of the referring expression.

2.3 Referring Expression Generation Process

As shown in Figure 3, the whole generation process is conducted in four stages. The prompts we use are provided in section 5.

Input Preparation We first filter the images to only keep those with at least 3 candidate instances. We then put a number tag on each instance, similar to Set-of-Marks (Yang et al., 2023), but since we already have instance annotations, we only need to add the number tag to the candidate instances.

2.3.1 LLM-Authored Pipeline

Before detailing the pipeline, we note an important design choice. We first attempted single step prompting of GPT-4o to directly produce complete referring expressions from the image and candidate instances. In practice, GPT-4o frequently produced overspecified descriptions with many redundant descriptors, which enabled shortcut grounding and weakened the need to understand the whole expression. To avoid this behavior, we adopt a two stage procedure: we first elicit discriminative attributes (between group A and group B and within group A), and then compose the final expression from a minimal yet sufficient subset of those attributes.

Similarity Judgement If there is a hard distractor and a target instance, they will be similar in some ways. To encourage the LLMs to identify any such similar pair in the image, we define two groups, group A and group B, where group A contains the hard distractor and the target instance, and group B contains the other distractors. We then prompt the LLMs to identify the two groups and to describe (1) attributes that distinguish the groups and (2) attributes that distinguish the two instances within group A. We ask for multiple alternative descriptions for each distinction. This could help us generate multiple diverse referring expressions for one image and allow us to select the high quality ones.

Referring Expression Generation After the similarity judgement, we obtain a list of paired descriptors that distinguish (1) group A from group B and (2) the two instances within group A. To ensure naturalness and diversity in phrasing, we prompt LLMs to compose referring expressions from combinations of these descriptors. Specifically, we use two alternative strategies: (1) employ the target's descriptors and (2) use the negation of the hard distractor's descriptors. This promotes more diverse and natural expressions. We also explicitly instruct the LLMs to not include number tag related descriptions. Although the elicited descriptors alone are sufficient for generation, we find that including the image input at this stage yields more diverse and accurate expressions, so

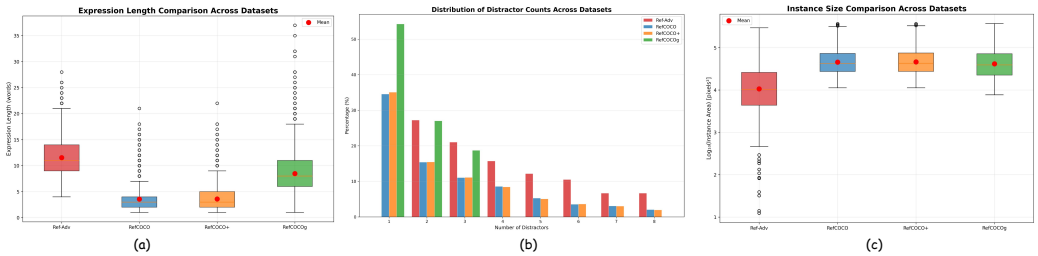


Figure 4: Dataset statistics across REC benchmarks. (a) Expression length comparison. (b) Distribution of distractor counts. (c) Instance size on a log area scale.

Table 2: Accuracy@0.5 after replacing the original referring expressions with the fixed "the one" prompt. Δ is Fixed@0.5 minus Ref-Adv Fixed@0.5 (shown in blue). With fixed prompt, models achieve higher accuracy on RefCOCO, RefCOCO+, and RefCOCOg than Ref-Adv.

Model	RefCOCO		RefCOCO+		RefCOCOg		Ref-Adv
	Fixed@0.5	Δ vs Ref-Adv	Fixed@0.5	Δ vs Ref-Adv	Fixed@0.5	Δ vs Ref-Adv	Fixed@0.5
Qwen2.5-VL-72B	35.1%	+13.7%	39.4%	+18.0%	38.3%	+16.9%	21.4%
InternVL-3-14B	35.9%	+13.1%	38.0%	+15.2%	38.2%	+15.4%	22.8%

we include the image. After this stage, we obtain multiple candidate referring expressions for each target instance.

2.3.2 Human-Authored Pipeline

We also collect a subset of human-authored referring expressions. For each filtered image, annotators first confirm whether there is a hard distractor pair and, if so, write a referring expression for it. Annotators are instructed to produce diverse and natural phrasing.

2.3.3 Verification Protocol

We verify each image–text pair. Three annotators answer two questions: (1) whether the expression is correct and unambiguous and (2) whether hard distractors are present in the image. Annotators first attempt grounding on the original image (without number tags) using the LLM generated expression. We then show the ground truth box overlaid on the image for reference, allowing reflection if their initial grounding was incorrect. Afterward, annotators record their final decisions on correctness/unambiguity and on the presence of hard distractors. Pairs are presented in a random order per annotator, and a pair is kept only if all three annotators agree. The keep rate is 18.7% for LLM-authored expressions.

2.4 Quality Analysis

Despite verification to ensure correctness, there remain potential issues for an REC benchmark that could affect fairness and the evaluation of reasoning skills. To further assess the quality of our data, we conduct the following analyses.

Statistics As shown in Figure 4 and table 1, Ref-Adv exhibits clear advantages in expression length, vocabulary size, distractor counts, and the negation ratio.

Model Bias Test Inspired by Cirik et al. (2018); Chen et al. (2020), we conduct a bias test of modern MLLMs (Qwen2.5-VL-72B and InternVL-3) on RefCOCO(+/g) and Ref-Adv. Here, bias refers to statistical regularities that may arise if training data comes from the same source as an evaluation benchmark, which can benefit performance. We design the test as follows: we replace the referring expression with a fixed prompt ("the one"), keep the same image, and prompt the model to output a bounding box. This test reveals whether model bias helps localize the target. The results are shown in table 2. They suggest that Ref-Adv is less affected by this bias than other benchmarks.

Table 3: Bag-of-words ablation on RefCOCO, RefCOCO+, RefCOCOg, and Ref-Adv. Acc@0.5 with original expressions vs bag-of-words (word order removed). Δ denotes (BoW − Original).

Model	RefCOCO			RefCOCO+			RefCOCOg			Ref-Adv		
	Orig@0.5	BoW@0.5	Δ	Orig@0.5	BoW@0.5	Δ	Orig@0.5	BoW@0.5	Δ	Orig@0.5	BoW@0.5	Δ
Qwen2.5-VL-72B	92.7%	82.8%	-9.9%	88.9%	78.2%	-10.7%	89.9%	75.3%	-14.6%	58.3%	41.5%	-16.8%
InternVL-3-14B	92.0%	84.7%	-7.3%	87.6%	81.0%	-6.6%	88.5%	74.9%	-13.6%	52.3%	38.6%	-13.7%

Table 4: One descriptor deletion ablation on RefCOCO, RefCOCO+, RefCOCOg, and Ref-Adv. Acc@0.5 with original expressions vs one descriptor deletion (removing a single descriptor in expression). Δ denotes (1-Desc − Original).

Model	RefCOCO			RefCOCO+			RefCOCOg			Ref-Adv		
	Orig@0.5	1D@0.5	Δ	Orig@0.5	1D@0.5	Δ	Orig@0.5	1D@0.5	Δ	Orig@0.5	1D@0.5	Δ
Qwen2.5-VL-72B	92.7%	88.0%	-4.7%	88.9%	83.6%	-5.3%	89.9%	85.3%	-4.6%	58.3%	51.9%	-6.4%
InternVL-3-14B	92.0%	87.1%	-4.9%	87.6%	82.4%	-5.2%	88.5%	83.8%	-4.7%	52.3%	45.2%	-7.1%

Textual Reasoning Necessity Test Prior work (Akula et al., 2020) shows that shuffling word order in RefCOCOg often leaves performance largely intact, indicating weak necessity for textual reasoning in prior REC benchmarks. This lack of degradation could stem from two factors: (1) expressions that only mention the target (or its parts) without referencing distractors and (2) images with no or very few distractors. Both factors reduce the reasoning demand in REC. To validate that Ref-Adv requires reasoning, we extend the test to RefCOCO(+/g) and Ref-Adv for comparison. Rather than shuffling while preserving meaning, we propose a simpler test: we convert the expression to a bag of words and randomize its order in the prompt (e.g., "a red ball with yellow stripes" becomes "with yellow red ball stripes a"). We evaluate Qwen2.5-VL-72B and InternVL-3 under this setting. Results are shown in table 3, indicating that Ref-Adv indeed requires texual understan and reasoning follow the referring expression exactly.

Avoidance of "Grounding Shortcut" As illustrated in Figure 1, RefCOCO(+/g) admits a "grounding shortcut," where a model can localize the target by checking a small subset of descriptors, without reasoning over the entire expression. To validate that Ref-Adv avoids this shortcut, we conduct a ***descriptor-deletion sufficiency*** test. For a given referring expression, we first use Qwen2.5-72B (Team, 2024) to extract all descriptors, randomly delete one, and ask Qwen2.5-72B to rewrite the expression with that descriptor removed. We then evaluate MLLMs on the modified image–text pair. If deleting a descriptor does not affect performance, the descriptor is unnecessary, suggesting a shortcut that succeeds without understanding the full expression. Such shortcuts are exacerbated in datasets with imbalanced numbers of descriptors and distractors. Results are shown in table 4, indicating that Ref-Adv has far fewer grounding shortcuts than others.

3 Experiment

3.1 Evaluation Setup

Evaluated Models We evaluate contemporary state of the art MLLMs, both closed source and open source, on Ref-Adv. The suite includes Qwen2.5-VL series (Cloud, 2025), InternVL-3 series (Laboratory, 2025), Gemini 2.5-Flash (Google, 2025), Gemini 2.5-Pro (Google, 2025), CogVLM-Grounding (THUDM, 2024), GLM-4.5V (ZhipuAI, 2025), GPT-4o (OpenAI, 2024), and Claude-3.5 Sonnet (Anthropic, 2024).

Evaluation Methods Set-of-Marks (SoM) overlays numbered marks on candidate objects in the image and leverages a specialized segmenter to provide fine-grained localization, avoiding the need for the MLLM to perform grounding itself. Because GPT-4o and Claude-3.5 have limited grounding ability, we evaluate them using SoM (Yang et al., 2023) with Semantic-SAM (Li et al., 2023). We use Semantic-SAM due to its strong performance on COCO images, one of the sources of Ref-Adv.

For each model (except CogVLM-Grounding which does not support CoT), we evaluate both with and without Chain-of-Thought (CoT). While CoT is uncommon in classic REC benchmark evaluation, Ref-Adv requires more reasoning, so we include CoT in our setup. Table 6 and table 7 report results on Ref-Adv and RefCOCO(+/g) with and without CoT.

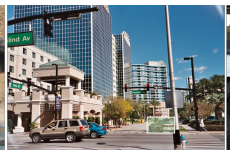
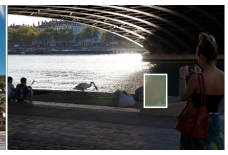
	LLM-Authored			Human-Authored
Image				
Expression	The cup that is empty and positioned at the center of the pizzas.	The young elephant near the water not covered in more dust.	The car is brown or black with a lighter tint on its windows.	The person sitting farther to the one holding a goose feather
Image				
Expression	The frisbee holder who is not facing forward.	The occupied chair with the person's arm not resting on the lap.	The sailboat, identified by being anchored, is the one with a flag.	The frisbee held by the person on the right hand side of the person in red

Table 5: Examples from Ref-Adv. Columns 1 to 3 are LLM generated; column 4 is human authored.

Table 6: Main results on Ref-Adv. Rows list models; columns report accuracy at IoU thresholds 0.5, 0.75, and 0.9, and mean accuracy (mAcc). We also report results for expressions with negation (explicit negation in the descriptor) and for varying numbers of distractors.

Model	Setting		Acc0.5	Acc0.75	Acc0.9	mAcc	Negation (Acc0.5)		Distractors (Acc0.5)		
	CoT?	SoM?					Neg	No-Neg	2–3	4–6	≥7
GPT-4o 2024	✗	✓	52.3	31.2	13.4	27.8	48.7	53.3	55.1	53.4	51.7
GPT-4o 2024	✓	✓	63.7	38.4	19.7	34.1	59.2	64.9	65.3	62.9	60.5
Claude-3.5 Sonnet 2024	✗	✓	40.8	22.1	3.8	22.4	37.5	41.7	43.6	39.0	37.4
Claude-3.5 Sonnet 2024	✓	✓	45.2	19.8	2.1	23.3	41.8	46.1	46.1	44.2	42.3
Gemini 2.5-Flash 2025	✗	✗	50.6	23.7	6.9	19.2	47.3	51.5	53.1	49.5	48.9
Gemini 2.5-Flash 2025	✓	✗	59.4	35.1	16.3	30.6	55.9	60.3	60.6	58.1	55.6
Gemini 2.5-Pro 2025	✗	✗	51.9	28.4	11.7	23.7	47.3	53.2	54.9	50.3	49.7
Gemini 2.5-Pro 2025	✓	✗	59.1	32.6	14.2	28.3	56.1	60.0	60.1	58.0	55.9
InternVL-3-7B 2025	✗	✗	49.5	39.2	21.4	33.1	47.1	50.1	51.8	49.2	48.6
InternVL-3-7B 2025	✓	✗	48.7	37.9	20.1	31.8	44.4	50.9	49.2	47.5	45.8
InternVL-3-14B 2025	✗	✗	50.5	40.7	22.8	34.2	48.9	51.2	51.1	49.7	50.3
InternVL-3-14B 2025	✓	✗	52.3	42.1	24.3	35.6	49.2	53.0	52.7	51.9	49.1
InternVL-3-38B 2025	✗	✗	53.8	43.5	25.7	37.1	50.1	54.8	55.9	53.4	52.9
InternVL-3-38B 2025	✓	✗	57.2	46.8	28.9	40.3	53.6	58.2	57.7	56.9	54.1
InternVL-3-78B 2025	✗	✗	54.6	44.2	26.4	37.8	51.9	55.6	57.4	53.9	53.4
InternVL-3-78B 2025	✓	✗	58.4	47.9	29.6	41.2	55.3	59.3	59.0	57.2	55.4
Qwen2.5-VL-7B 2025	✗	✗	49.3	39.0	21.2	32.9	46.9	50.0	50.9	48.4	48.1
Qwen2.5-VL-7B 2025	✓	✗	49.1	38.8	20.9	32.7	45.7	50.0	50.2	47.6	46.0
Qwen2.5-VL-32B 2025	✗	✗	52.7	42.4	24.6	36.0	48.4	53.6	55.0	52.5	52.0
Qwen2.5-VL-32B 2025	✓	✗	56.8	46.5	28.7	40.1	53.9	57.6	57.3	55.8	54.3
Qwen2.5-VL-72B 2025	✗	✗	54.1	43.8	25.9	37.4	50.9	55.3	56.6	54.1	53.6
Qwen2.5-VL-72B 2025	✓	✗	58.3	47.8	29.5	41.1	55.2	59.4	58.6	58.1	55.6
GLM-4.5V 2025	✗	✗	52.4	42.1	24.3	35.6	49.3	53.1	54.2	51.9	51.6
GLM-4.5V 2025	✓	✗	56.9	46.6	28.8	40.2	53.9	57.7	57.2	55.9	54.6
CogVLM-Grounding 2024	✗	✗	51.5	41.2	23.4	35.0	49.2	52.2	54.0	52.4	50.8

Evaluation Prompts Models differ in prompt format and output conventions. For example, Qwen2.5-VL-72B uses absolute coordinates, while others use normalized coordinates; CogVLM-Grounding requires the question to strictly follow the form "Where is the 'referring expression'?" to output boxes. To ensure fairness, we adopt best-practice prompts for each model.

3.2 Evaluation Metrics

Accuracy serves as a widely adopted metric for evaluating existing REC models. A referring expression instance is deemed successfully grounded when the Intersection over Union (IoU) between the predicted bounding box and the ground truth annotation surpasses 0.5. This conventional evaluation

Table 7: RefCOCO(+/g) and Ref-Adv Acc@0.5 with and without Chain-of-Thought (CoT). Δ denotes (CoT − Direct).

Model	RefCOCO			RefCOCO+			RefCOCOg			Ref-Adv		
	Direct	CoT	Δ	Direct	CoT	Δ	Direct	CoT	Δ	Direct	CoT	Δ
Qwen2.5-VL-72B	92.7%	89.3%	-3.4%	88.9%	86.2%	-2.7%	89.9%	88.5%	-1.4%	54.1%	58.3%	+4.2%
InternVL-3-14B	92.0%	89.2%	-2.8%	87.4%	85.8%	-1.6%	88.6%	87.4%	-1.2%	50.5%	52.3%	+1.8%

Figure 5: Performance of representative multimodal LLMs on Ref-Adv. We include qualitative examples with and without CoT for Gemini 2.5-Flash and Qwen2.5-VL-72B. CoT answers are shown in a gray box. Hard distractors in Ref-Adv challenge current MLLMs.

metric is designated as Acc0.5. Here, we implement multiple evaluation protocols, i.e., Accuracy computed under different IoU thresholds such as Acc0.5, Acc0.75, Acc0.9, and mean Accuracy (mAcc) across all IoU criteria, to thoroughly evaluate the precision and robustness.

3.3 Analysis

Effect of CoT Table 6 and table 7 show that CoT generally improves performance on Ref-Adv, while it can reduce accuracy on RefCOCO(+/g). We attribute the improvement on Ref-Adv to its heavier reasoning demand; for RefCOCO(+/g), where grounding can often succeed without extensive reasoning, CoT may introduce unnecessary verbosity or error.

Main Results Table 6 summarizes results on Ref-Adv. With SoM, GPT-4o attains the best performance on Ref-Adv under CoT, suggesting strong reasoning and visual perception capabilities. While other models perform well on RefCOCO(+/g), their accuracy drops markedly on Ref-Adv, revealing gaps in visual reasoning and perception.

Qualitative Analysis Figure 5 shows qualitative examples for Qwen2.5-VL-72B and Gemini 2.5-Flash, both with and without CoT. With explicit reasoning, models often follow the intended chain, but in harder cases they fail partway due to incorrect visual perception or a misunderstanding of the referring expression. Notably, models often select the hard distractor as the answer, which indicates that Ref-Adv challenges models to both deeply understand referring expressions and perform

accurate visual perception. This suggests that Ref-Adv stresses advanced reasoning and visual perception, and that current state of the art MLLMs still exhibit clear gaps.

4 Literature Review

Referring Expression Benchmarks. The field's foundational benchmarks, including the ReferItGame (Kazemzadeh et al., 2014) and the de facto standard RefCOCO suite (RefCOCO/+/g) (Yu et al., 2016; Mao et al., 2016), have been instrumental in advancing research. However, subsequent analyses revealed that high scores on these datasets can overstate genuine grounding abilities. For example, performance on RefCOCOg often remains high even with shuffled word order, indicating a reliance on superficial cues rather than robust compositional understanding (Akula et al., 2020). To address these cracks in the foundation—namely simplistic expressions and a lack of hard, same-category distractors—a new wave of benchmarks emerged. To directly target reasoning, Cops-Ref (Chen et al., 2020) and its successor FineCops-Ref (Liu et al., 2024) introduced more compositional expressions with explicit distractors and negative examples, while the synthetic CLEVR-Ref+ (Liu et al., 2019) offered a fully controlled environment for diagnostic analysis. Concurrently, other efforts expanded the scope of the REC task itself. gRefCOCO (Liu et al., 2023) introduced multi-target and no-target expressions, PhraseCut (Wu et al., 2020) scaled up to phrase-level segmentation over more categories, and recent works like HC-RefLoCo (Wei et al., 2024) and Ref-L4 (Chen et al., 2024) have pushed for longer, more natural descriptions and corrected label noise in the original benchmarks.

The need for such challenging benchmarks is further amplified by the rapid advancements in Multimodal Large Language Models (MLLMs), which now dominate the field.

Multimodal Large Language Models. Recent progress in vision language AI has been driven by large multimodal language models (MLLMs) that combine powerful LLM backbones with vision encoders and alignment tuning for instruction following. Proprietary models like OpenAI's GPT-4 Vision and Google's Gemini exemplify this trend, while open source counterparts such as Alibaba's Qwen-VL and Shanghai AI Lab's InternVL offer similar capabilities (OpenAI, 2024; Google, 2025; Cloud, 2025; Laboratory, 2025). These systems, trained on massive image text corpora, now achieve near ceiling accuracy (often $>90\%$) on classic referring expression benchmarks (Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016). However, as the reasoning capabilities of MLLMs rapidly advance, it has become clear that these high scores are insufficient to measure genuine multi-step reasoning, necessitating an evolution in the REC task itself (Wei et al., 2024; Chen et al., 2024). This has spurred the development of both more challenging benchmarks and reasoning enhanced models. For example, Moonshot's Kimi-VL (Thinking) applies chain of thought fine tuning and reinforcement learning to strengthen stepwise visual reasoning (AI, 2025), and ZhipuAI's GLM-4.5V explicitly performs step by step grounding to output precise object bounding boxes (ZhipuAI, 2025). Similarly, new aligned vision language models like CogVLM and DeepSeek-VL2 incorporate mixture of experts or reward optimization to improve visual grounding and coherence, and even commercial chatbots (e.g., Anthropic's Claude 3.5, xAI's Grok) are beginning to integrate advanced multimodal reasoning. Our work builds on these efforts by evaluating a broad suite of state of the art MLLMs—both general purpose and reasoning centric—on a novel REC benchmark designed to stress test their visual grounding and reasoning abilities (THUDM, 2024; ZhipuAI, 2025; AI, 2025; DeepSeek, 2024; Anthropic, 2024; xAI, 2025).

5 Conclusion

In this work, we introduced Ref-Adv, a modern REC benchmark designed to address the reliance on visual shortcuts in existing datasets by requiring genuine multi-step reasoning. We construct Ref-Adv through a two stage pipeline that use an LLM to compose minimally sufficient referring expressions. Our comprehensive ablation studies (section 2) confirm that Ref-Adv effectively probes both complex textual and visual grounding capabilities. Strikingly, our evaluation of contemporary MLLMs (section 3) revealed a significant performance drop compared to their near-saturated scores on RefCOCO(+/g), exposing a critical overestimation of their visual reasoning abilities. These findings underscore the urgent need for benchmarks that reflect real-world visual complexity and offer a clear path forward for developing more robust and capable MLLMs.

Ethics Statement

We follow the ICLR Code of Ethics (https://iclr.cc/public/CodeOfEthics). We use large language models to draft candidate expressions and then apply a human verification step with three annotators to ensure correctness and remove ambiguous or unsafe content (section 2). Annotators worked only with public images and could skip any example. Our benchmark is intended for evaluating grounding and visual reasoning, not for surveillance or biometric identification. We release only expressions, target regions, and dataset identifiers, and we provide usage guidance that discourages applications involving identity inference or sensitive attribute prediction. We are not aware of conflicts of interest.

Reproducibility Statement

Section 2 describes the complete data pipeline, including image sources, filtering with same-class distractors, descriptor elicitation, expression composition, and the three-annotator verification protocol, with a step-by-step diagram in Figure 3. We will release the exact image identifiers, the final referring expressions, target regions, and the JSON schema of our annotations, together with scripts to load and evaluate the data. Evaluation protocols and metrics (Acc0.5/Acc0.75/Acc0.9 and mean Accuracy) are specified in Section 3. To facilitate exact replication, we will provide below artifacts upon publication: (i) the evaluation scripts that compute IoU and accuracy, (ii) the prompts and configuration files for each evaluated model. Together, these artifacts enable end-to-end reproduction of our tables and figures.

References

Moonshot AI. Kimi vl (thinking). https://kimi.moonshot.cn, 2025.

Arjun R. Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6555–6565, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.586. URL https://aclanthology.org/2020.acl-main.586/.

Anthropic. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S.-H. Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *arXiv preprint arXiv:2406.16866*, 2024. URL https://arxiv.org/abs/2406.16866.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10086–10095, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_Cops-Ref_A_New_Dataset_and_Task_on_Compositional_Referring_Expression_CVPR_2020_paper.html.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. Visual referring expression recognition: What do systems actually learn? *arXiv preprint arXiv:1805.11818*, 2018.

Alibaba Cloud. Qwen 2.5 vl. urlhttps://qwen.vl.alibabacloud.com/, 2025.

DeepSeek. Deepseek-vl2. https://github.com/deepseek-ai/DeepSeek-VL, 2024.

Google. Gemini 2.5 flash. urlhttps://deepmind.google/technologies/gemini/flash/, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

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

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Shanghai AI Laboratory. Internvl 3.
urlhttps://internvl.github.io/, 2025.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. URL https://arxiv.org/abs/2307.04767.

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

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Liu_GRES_Generalized_Referring_Expression_Segmentation_CVPR_2023_paper.html.

Junzhuo Liu, Xuzheng Yang, Weiwei Li, and Peng Wang. Finecops-ref: A new dataset and task for fine-grained compositional referring expression comprehension. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15440–15457, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.864. URL https://aclanthology.org/2024.emnlp-main.864/.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4185–4194, June 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Liu_CLEVR-Ref_Diagnosing_Visual_Reasoning_With_Referring_Expressions_CVPR_2019_paper.html.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/papers/Mao_Generation_and_Comprehension_CVPR_2016_paper.pdf.

OpenAI. Gpt-4o. https://openai.com/index/gpt-4o/, 2024.

Qwen Team. Qwen2.5: Large language and vision-language models.
urlhttps://qwen.readthedocs.io/, 2024. Technical Report.

THUDM. Cogvlm2-grounding. https://github.com/THUDM/CogVLM2, 2024.

Fangyun Wei, Jinjing Zhao, Kun Yan, Hongyang Zhang, and Chang Xu. A large-scale human-centric benchmark for referring expression comprehension in the lmm era. In *NeurIPS Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/80f0cd0305f7741659304f5325f3bf6d-Paper-Datasets_and_Benchmarks_Track.pdf.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. URL https://arxiv.org/abs/2201.11903.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10216–10225, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wu_PhraseCut_Language-Based_Image_Segmentation_in_the_Wild_CVPR_2020_paper.html.

xAI. Grok-4 fast. https://x.ai, 2025.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. URL https://arxiv.org/abs/2310.11441.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pp. 69–85. Springer, 2016. doi: 10.1007/978-3-319-46475-6_5. URL https://link.springer.com/chapter/10.1007/978-3-319-46475-6_5.

ZhipuAI. Glm-4.5v. https://www.zhipuai.cn/en, 2025.

A USE OF LLM IN WRITING.

We employed large language models (LLMs) to assist in polishing the text throughout this paper, including refining phrasing, improving clarity, and ensuring grammatical correctness.

B PROMPT IN DATA COLLECTION

We include the core prompt templates used by our two-stage LLM-authored pipeline described in section 2. Query 1 elicits group-level and intra-pair discriminators; Query 2 composes minimally sufficient referring expressions from those discriminators. Placeholders such as {num_objects} and {target_class} are filled at runtime.

We use structured output in JSON format for the LLMs to ensure the output is in the correct format.

```
You are given an image with {num_objects} {target_class} objects labeled
    by integers (1..N).

**Task**:
1) Choose the most similar pair `{{i,j}}` and call that group **A**.
    Everything else is group **B**.
2) Propose exactly **2 group-level discriminators** to separate **A vs B
    **. Each discriminator must have an A-side phrase and a B-side phrase
    .
3) For the two {target_class} objects inside A, propose exactly **4 intra
    -pair discriminators** (2 "noticeable", 2 "unnoticeable"). Each must
    provide a phrase for object `i` and a phrase for object `j`, plus a "
    noticeability" field with value "noticeable" or "unnoticeable".

**Output JSON only**, matching this schema (no extra text):
{{
  "similar_group": {{"ids":[int,int], "label":"A"}},
  "groups": {{"A":[int,...], "B":[int,...]}},
  "group_discriminators":[
    {{"id":"G1","name":string,"A":string,"B":string}},
    {{"id":"G2","name":string,"A":string,"B":string}}
  ],
  "in_pair_discriminators":[
    {{"id":"P1","name":string,"i":string,"j":string,"noticeability":"
        noticeable or unnoticeable"}},
    {{"id":"P2","name":string,"i":string,"j":string,"noticeability":"
        noticeable or unnoticeable"}},
    {{"id":"P3","name":string,"i":string,"j":string,"noticeability":"
        noticeable or unnoticeable"}},
    {{"id":"P4","name":string,"i":string,"j":string,"noticeability":"
        noticeable or unnoticeable"}}
  ]
}}

If the model is multimodal, attend to the image; otherwise rely on the
    provided description/annotations.
```

Listing 1: Query 1: Similarity Judgement and Discriminator Elicitation

C LLM API COST FOR DATA COLLECTION

The kept rate is 18.7% for a LLM-authored expression, and each expression will cost about 2300 input tokens and 120 output tokens, with GPT-4o price of \$2.5 per 1M input tokens and \$10 per 1M output tokens, the cost for a LLM-authored expression is $(2300 \times 2.5 + 120 \times 10)/1,000,000 = \$0.00695$. Given that we need to generate approximately $1/0.187 = 5.35$ expressions to get one

```
System: You are a visual assistant that returns JSON only. Follow the
    user's schema exactly. Do not include any extra text.

Image context template: This is an image with {num_objects} {target_class
    }(s) overlaid with integers (1..N).

{image_context}

You are given some observations and a `target_id`.

**Observations**:
{query1_json}

**Target ID**: {target_id}
**Target Class**: {target_class}

**Task**: Write the referring expressions that refer to {target_class} `
    target_id` based on the observations. Each sentence should use one
    group discriminator (A vs B) and one intra-pair discriminator (
    between the two in A). Return 4 in total.

Return JSON only with this schema:
{{
  "expressions": [
    {{"id":"E1","target_id":int,"group_dids":["G?"],"pair_dids":["P?"],"
        inpair_positive_phrase":string,"inpair_negative_phrase":string,"
        inpair_phrase":"only_positive|only_negative|both","text":string
        }},
    {{"id":"E2","target_id":int,"group_dids":["G?"],"pair_dids":["P?"],"
        inpair_positive_phrase":string,"inpair_negative_phrase":string,"
        inpair_phrase":"only_positive|only_negative|both","text":string
        }},
    {{"id":"E3","target_id":int,"group_dids":["G?"],"pair_dids":["P?"],"
        inpair_positive_phrase":string,"inpair_negative_phrase":string,"
        inpair_phrase":"only_positive|only_negative|both","text":string
        }},
    {{"id":"E4","target_id":int,"group_dids":["G?"],"pair_dids":["P?"],"
        inpair_positive_phrase":string,"inpair_negative_phrase":string,"
        inpair_phrase":"only_positive|only_negative|both","text":string}}
  ]
}}

Explanation example for `inpair_phrase`: if `inpair_positive_phrase` is "
    sitting" and `inpair_negative_phrase` is "standing", then "
    only_positive" means "the one sitting"; "only_negative" means "the
    one not standing"; "both" means "the one sitting rather than standing
    ".

Constraints: Use different combinations of group_dids and pair_dids. Vary
     phrasings and sentence structures. Do not mention numeric labels in
    the text.
```

Listing 2: Query 2: Referring Expression Composition

kept expression, the effective cost per kept expression is $5.35 \times \$0.00695 = \$0.0372$. For our dataset of 4,000 expressions, the total cost is approximately $4000 \times 0.0372 = \$148.8$.