This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

GroundingME: Exposing the Visual Grounding Gap in MLLMs through Multi-Dimensional Evaluation

Rang Li[1,*] Lei Li[2,3] Shuhuai Ren[2] Hao Tian[2] Shuhao Gu[2] Shicheng Li[1,2] Zihao Yue[2,4]
Yudong Wang[1,2] Wenhan Ma[1,2] Zhe Yang[1] Jingyuan Ma[1] Zhifang Sui[1,◇] Fuli Luo[2,◇]

[1]State Key Laboratory of Multimedia Information Processing,
School of Computer Science, Peking University
[2]LLM-Core Xiaomi [3]The University of Hong Kong [4]Renmin University of China
lirang410@gmail.com, szf@pku.edu.cn, luofuli@xiaomi.com

Abstract

Visual grounding, localizing objects from natural language descriptions, represents a critical bridge between language and vision understanding. While multimodal large language models (MLLMs) achieve impressive scores on existing benchmarks, a fundamental question remains: can MLLMs truly visually ground with human-like sophistication, or are they merely pattern-matching on simplified datasets? Current benchmarks fail to capture real-world complexity where humans effortlessly navigate intricate references and recognize when grounding is impossible. To rigorously assess MLLMs' true capabilities, we introduce GroundingME, a benchmark that systematically challenges models across four critical dimensions: (1) Discriminative: distinguishing highly similar objects, (2) Spatial: understanding complex relational descriptions, (3) Limited: handling occlusions or tiny objects, and (4) Rejection: recognizing ungroundable queries. Through careful curation combining automated generation with human verification, we create 1,005 challenging examples mirroring real-world complexity. Evaluating 25 state-of-the-art MLLMs reveals a profound capability gap: the best model achieves only 45.1% accuracy, while most score 0% on rejection tasks. We explore two strategies for improvements: (1) test-time scaling selects optimal response by thinking trajectory to improve overall performance by up to 4.5%, and (2) data-mixture training boosts rejection accuracy from 0% to 27.9%. GroundingME thus serves as both a diagnostic tool revealing current limitations in MLLMs and a roadmap toward human-level visual grounding. Project page: https://groundingme.github.io.



Figure 1. **Examples of different visual grounding benchmarks.** Prior benchmarks (Top) are either too simple or prone to shortcuts. Our proposed GroundingME (Bottom) increases the challenge in four important dimensions. **The green bounding box** indicates the correct ground-truth object, while **the red bounding box** shows the answer of Qwen3-VL-30B-A3B-Instruct. Critical information for the answer is highlighted in the description.

1. Introduction

The rise of Multimodal Large Language Models (MLLMs) represents a paradigm shift in artificial intelligence, offering unprecedented capabilities in joint vision and language understanding [4, 12, 20, 25]. Visual grounding [26, 42],

*Work done during internship at Xiaomi Corporation.
◇Co-corresponding authors.

the task of localizing a specific region in an image based on a natural language description, also known as Referring Expression Comprehension (REC) [16, 24, 28], stands as a fundamental capability for these models. It is the bedrock for enabling complex, language-driven interactions, facilitating precise, real-world applications from robotic instruction [37, 50] to detailed image editing [6, 31].

However, the remarkable reasoning abilities demonstrated by MLLMs are largely untested in complex grounding scenarios. As shown in Fig. 1, early benchmarks [15, 23, 48] are fundamentally limited to simple phrases or basic spatial relations in uncluttered scenes (e.g., "the vase on the right"), failing to assess the fine-grained appearance discrimination, language understanding, and complex spatial reasoning that modern MLLMs claim to possess. While recent works [8, 40] have attempted to increase task difficulty with longer descriptions, they often fail to increase the actual reasoning complexity, as models can easily take a shortcut by relying on simple keyword matching (e.g., a unique class name) while bypassing the complex attribute and spatial information. As a result, recent models have already achieved over 90% accuracy [4, 34, 39, 43] on the RefCOCO series [23, 48] and nearly 90% accuracy [34] on Ref-L4 [8], indicating that existing benchmarks can no longer differentiate the true grounding abilities of current models. Additionally, these works overlook the model ability to reject a description when its fine-grained details do not precisely match the visual evidence, which is critical for safety and reliability in real-world applications.

To bridge this gap, we introduce GroundingME, a comprehensive visual grounding benchmark comprising 1,005 samples, specifically designed for the rigorous evaluation of MLLMs. The benchmark is constructed via a three-stage process: (1) Bounding Box Annotation; (2) Description Generation; and (3) Manual Selection and Refinement. We design a challenge taxonomy that systematically evaluates models across four L-1 dimensions, as shown in Fig. 1. This taxonomy comprehensively tests model performance across distinct challenge types: (1) Discriminative focuses on distinguishing objects based on subtle, fine-grained visual differences; (2) Spatial assesses the ability to understand and resolve complex spatial and relational arrangements; (3) Limited evaluates grounding under conditions of minimal visual features due to external constraints; and (4) Rejection tests the ability to reject a misleading description that contains subtle errors. Furthermore, we provide a fine-grained L-2 hierarchy covering twelve subcategories to enable a deeper, diagnostic analysis of model performance.

We conduct an extensive evaluation across 25 state-of-the-art commercial and open-source models, including the Qwen3-VL series [4], Gemini-2.5 [9], Seed-1.6-Vision [13], and GLM-4.5V [34], with parameter sizes ranging from 2B to 235B. The results reveal their widespread and significant shortcomings on our benchmark. Even the top-tier model, Qwen3-VL-235B-A22B, achieves only 45.1% accuracy, with the majority of models scoring between 10% and 40%. The failure is most severe on the Rejection category, with most models scoring 0%. We find this failure persists even as model scale increases.

Table 1. **Comparison between GroundingME and other visual grounding benchmarks or datasets.** Comparison aspects include: General Scenario and Object Type (General), Description with Compositional Semantics (Semantic), Multiple Evaluation Dimension (Multi-Dim.), and Rejection Samples (Rejection).

Benchmarks	General	Semantic	Multi-Dim.	Rejection
Refcoco/+/g [23, 48]	✔	✘	✘	✘
CLEVR-Ref+ [21]	✘	✔	✘	✘
Refcrowd [29]	✘	✘	✘	✘
Ref-ZOM [15]	✔	✘	✘	✔
HC-RefLoCo [40]	✘	✔	✘	✘
Ref-L4 [8]	✔	✔	✘	✘
Ref-Adv [10]	✔	✔	✘	✘
GroundingME	✔	✔	✔	✔

These widespread failures motivate us to explore strategies to improve model performance. We explore two complementary approaches: (1) at test time and (2) at training time. First, at test time, we observe that enabling thinking generally improves performance and enables basic rejection behavior. Building on this, we propose a Test-Time Scaling method [7, 32, 41], leveraging a judge model to select the optimal thinking quality from multiple candidates. Our experiments show this significantly improves performance across all subtasks, particularly on reasoning-intensive categories. Second, at training time, we hypothesize that the models' inability to reject stems from a lack of negative samples in training data. We test this with a simple Data-Mixture Training strategy. By fine-tuning Qwen3-VL-8B-Instruct [4] on RefCOCOg [23] augmented with negative samples, the model learns a foundational rejection capability, boosting its performance on our benchmark's Rejection category from 0% to 27.9%. Our findings reveal the limitations of current MLLMs and provide a clear and practical path forward for building more trustworthy visual systems.

2. Related Work

Multimodal Large Language Models (MLLMs). MLLMs have demonstrated remarkable capabilities in understanding and reasoning across both visual and textual data, pushing the boundaries of various multimodal tasks ,including visual grounding. Recent models [4, 34, 39, 43] have already achieved impressive performance on existing visual grounding benchmarks [8, 23, 48], with accuracy approaching or exceeding 90%. Moreover, these models

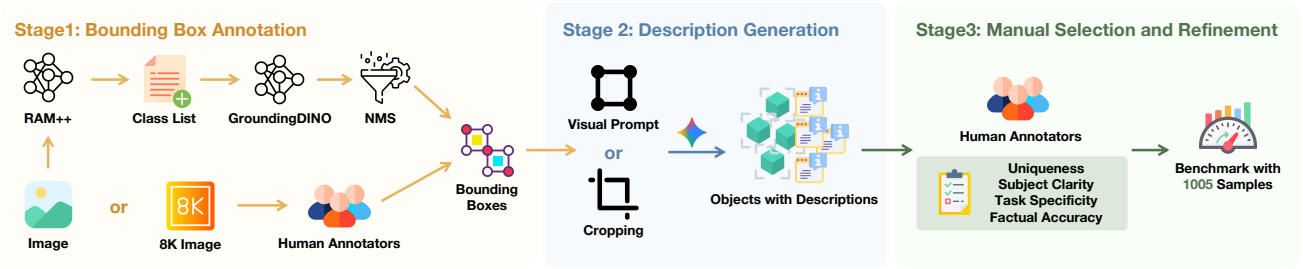


Figure 2. **The overall data construction pipeline of GroundingME.** The process consists of three main stages: (1) **Bounding Box Annotation**, which utilizes a semi-automated pipeline with RAM++ and GroundingDINO for bounding box generation (§3.2.1); (2) **Description Generation**, which leverages Gemini-2.5-Flash for generating initial referring expressions (§3.2.2); and (3) **Manual Selection and Refinement**, where human annotators apply rigorous filtering and refinement according to our challenge taxonomy (§3.2.3).

are capable of performing complex reasoning over visual contents to further enhance multimodal performance. This advancement creates a pressing need for more challenging and comprehensive benchmarks, as traditional datasets are often insufficient to rigorously evaluate their nuanced understanding and reasoning capabilities on complex, real-world tasks [42], motivating us to curate a more challenging visual grounding benchmark.

Visual Grounding Datasets and Benchmarks. The evaluation of visual grounding has evolved significantly over the years [10, 16, 24, 26, 28, 42], progressing from closed set, single objects, brief phrases to open vocabulary, generalized targets, and complex descriptions. Early benchmarks, such as the RefCOCO series [23, 48], provided a foundational testbed but were characterized by simple, short phrases and have now become largely saturated. In response to these limitations, subsequent works introduced datasets targeting specific challenges. For instance, CLEVR-Ref+ [21] was developed to diagnose compositional reasoning in a synthetic environment, while RefCrowd [29] focused on fine-grained discrimination in crowded scenes. Ref-ZOM [15] was the first to introduce simple negative samples to test model's rejection ability. More recently, benchmarks have been designed specifically for MLLMs, such as HC-RefLoCo [40] and Ref-L4 [8], which employ long, descriptive sentences as queries to achieve increased difficulty. While existing benchmarks have progressively addressed specific shortcomings, they still fall short of providing a sufficient challenge. To this end, as summarized in Tab. 1, our proposed benchmark is the first to unify these critical dimensions to comprehensively assess the advanced visual grounding capabilities of modern MLLMs.

3. GroundingME

In this section, we introduce the detailed construction process of GroundingME. We first outline the data source (§3.1), followed by a three-stage human-in-the-loop annotation pipeline (§3.2). Finally, we present the distribution and the statistics of GroundingME (§3.3).

3.1. Data Source

To ensure that GroundingME provides a challenging and complex evaluation, we meticulously curate its image pool by leveraging two high-quality, high-resolution source datasets: SA-1B [17] and HR-Bench [38]. We selected these two datasets as our image sources because they inherently meet our requirements for visual complexity and scale. The SA-1B dataset, which is widely used [19, 30], offers extensive resources of complex scenes and high object density, with 11 million images and 1.1 billion masks. HR-Bench offers ultra-high resolution with its 8K subset essential for creating tasks where minute objects are clearly resolvable. Therefore, we use HR-Bench as the source of the Small subcategory, and rely on SA-1B as the source for all other subtasks. For both datasets, we only use the raw, original images as input for our construction pipeline, without any pre-existing masks, QA pairs, or other annotations. This ensures that even if models encountered the source images during training, the task itself remains novel, thus effectively mitigating the risk of data contamination.

3.2. Data Construction

The construction of our benchmark is a multi-stage process designed to guarantee the quality and diversity of the resulting dataset. Our pipeline consists of three main stages: (1) Bounding Box Annotation, (2) Description Generation, and (3) Manual Selection and Refinement.

3.2.1. Bounding Box Annotation

For images sourced from SA-1B and HR-Bench, we employ distinct methodologies for bounding box generation. For images from SA-1B, we develop an automated pipeline that combines RAM++ [51], GroundingDINO [22], and a customized Non-Maximum Suppression (NMS) rule. In contrast, for HR-Bench images, we leverage a manual annotation due to the challenges posed by ultra-high resolution.

Our automated pipeline for SA-1B images comprises three main steps. (1) We first utilize RAM++ to identify all object categories within each image, generating a comprehensive list of class names. (2) Based on this list, we then format a text query for the GroundingDINO model, which is used to generate a series of bounding boxes for each image. To optimize the generation, we adjusted the model's filtering threshold. This step outputs a series of bounding boxes and their highest-similarity tokens, and we use the word which the highest-similarity token belongs to as the class name for each box. (3) Finally, to eliminate redundant bounding boxes, we apply a customized NMS rule. Instead of prioritizing boxes by area, our NMS strategy favors those belonging to classes with a higher instance count, yielding the final set of bounding boxes for each image.

3.2.2. Description Generation

Leveraging the powerful image understanding capabilities of modern MLLMs, we use Gemini-2.5-Flash [9] to generate preliminary descriptions for each bounding box, which serve as a foundation for our referring expressions. For objects in the SA-1B dataset, we utilize the model's visual prompting capability by framing the objects in the full-size image with a red bounding box and prompting the model to generate a description that includes both their visual attributes and spatial relationships. In contrast, for objects in the HR-Bench dataset, the bounding box regions are too small to be effectively prompted within the full image context. Therefore, we crop the bounding box regions and input them to the model, prompting it to generate descriptions of the objects' visual attributes only.

3.2.3. Manual Selection and Refinement

The final stage of our pipeline is a meticulous manual selection and refinement process, a crucial step designed to mitigate the inherent limitations of automated data generation. Our human annotators directly address potential inaccuracies in bounding boxes, hallucinations in descriptions, and the presence of redundant or simplistic examples.

We apply a rigorous set of filtering and selection criteria to ensure our benchmark's quality and challenge. To prevent models from completing the task by relying solely on class names, we filter out simple samples by removing any classes with fewer than three instances and objects with bounding boxes occupying more than 50% of the image. We then meticulously select the majority of our samples from object classes with an instance count higher than 5. We further enrich the benchmark's diversity by supplementing specific subtasks based on predefined rules, for instance, by selecting samples for Counting and Partial subcategories from scenes with eight or more instances, or for Text subcategory from descriptions containing numbers or strings.

The selected samples undergo a detailed refinement process to enhance their quality. Annotators first correct any inaccuracies in the bounding boxes. Crucially, they then modify the automatically generated descriptions to meet four key criteria: (1) Uniqueness, ensuring each description refers to one and only one object, or no object for Rejection samples; (2) Subject Clarity, explicitly identifying the target object, which is essential for complex Spatial samples; (3) Task Specificity, tailoring the description to match the assigned sub-task (e.g., adding ordinal words for Counting tasks); (4) Factual Accuracy, correcting any hallucinations or intentionally introduce factual errors for Rejection samples. Inter-annotator agreement on 50 random samples from the final dataset yields pairwise Cohen's kappa scores of 0.64–0.73 (avg. 0.69), indicating substantial consistency.

3.3. Data Analysis

3.3.1. Subtask Distribution

GroundingME incorporates a two-tiered classification system, consisting of four L-1 categories and twelve L-2 subcategories, designed to comprehensively assess model performance across various subtasks. The benchmark comprises a total of 1,005 samples, distributed across four L-1 subtasks: Discriminative (204, 20.3%), Spatial (300, 29.9%), Limited (300, 29.9%), and Rejection (201, 20.0%). Each L-1 category is further divided into L-2 subcategories. Both the Discriminative and Rejection category are composed of four L-2 subcategories: Appearance, Component, Text, and State, with approximately 50 samples allocated to each. The Spatial category is equally split between Relationship and Counting subcategories, and the Limited category is equally divided between Occlusion and Small subcategories. This balanced distribution ensures robust evaluation across all facets of visual grounding challenge.

3.3.2. Statistics Analysis

Tab. 2 presents key statistics for GroundingME, substantiating its challenging nature through several metrics: (1) **Object Class and Quantity.** The benchmark encompasses 241 distinct object classes, ensuring a broad coverage of real-world scenarios. The challenge of intra-class confusion is quantified by the high Intra-Class Count Quartile of (5, 7, 12), indicating a large number of similar distracting objects in the image. (2) **Image and Instance Size.** The image size (square root of area) ranges from 1,500 to 7,680, representing a magnitude increase compared to 83 - 610 in RefCOCO series and 30 - 3,767 in Ref-L4 [8]. The instance size (square root of area) ranges from 21 to 946, demonstrating a wide coverage of various object scales. Furthermore, the Instance Area Ratio (the area of an instance's bounding box divided by the image area) Quartile measures only (0.16%, 1.0%, 2.7%), indicating that the instances are notably small relative to the image, which significantly increases the task difficulty. (3) **Description Complexity.** The descriptions in the benchmark all consist of complete sentences or para-

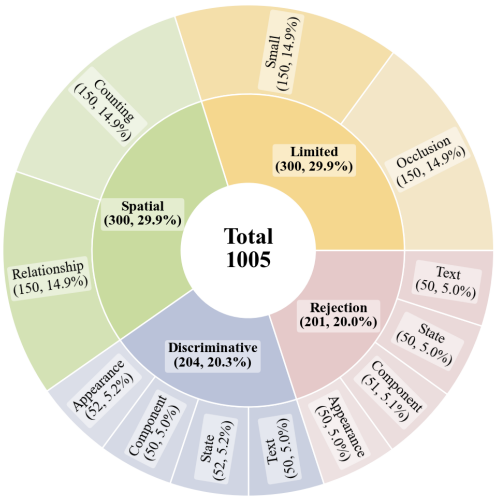


Figure 3. **Subtask Distribution of GroundingME.** Our benchmark comprises of 1,005 samples, distributed across four L-1 categories and twelve L-2 subcategories.

Table 2. **Statistics of GroundingME.** The image size and instance size are reported as the square root of their area in pixels. The description length is measured by the number of words. The Intra-Class Count Quartile is derived from the automated construction pipeline excluding images from HR-Bench, and thus should be interpreted as a lower bound for the actual value.

Statistics	Number
Object Class	241
Image Size	1,500 - 7,680
Instance Size	21 - 946
Instance Area Ratio Quartile	(0.16%, 1.0%, 2.7%)
Description Length Quartile	(18, 40, 58)
Intra-Class Count Quartile	(5, 7, 12)

graphs, making it suitable for modern MLLMs. The Description Length Quartile reaches (18, 40, 58) words, highlighting that the majority of descriptions are longer and more intricate than the average value of 3.6 in RefCOCO/+, 8.4 in RefCOCOg, and 24.2 in Ref-L4 [8].

4. Evaluation

We report our experimental setup and findings from the extensive evaluation of state-of-the-art MLLMs. This section is organized by models evaluated (§4.1), detailed evaluation settings (§4.2), and results analysis (§4.3).

4.1. Evaluated Models

We conduct a comprehensive evaluation of 25 state-of-the-art MLLMs on GroundingME, encompassing both open-source and commercial models. Our selection of open-source models is highly diverse, spanning 12 major publishers and a wide range of parameter sizes, including Qwen3-VL (2B/4B/8B/32B/A3B/A22B), Qwen2.5-VL (7B/32B/72B) [4], GLM-4.5V [34], InternVL3.5 (8B/A28B) [39], MiMo-VL-7B-RL-2508 [43], Keye-VL-1.5-8B [35], MiniCPM-V-4.5 [49], Phi-4-Multimodal [1], Llama-4 (Maverick/Scout) [11], LLaVA-OneVision-1.5-8B [3], Mistral-3.2-24B [36], Gemma-3-27B [33], and Llama-Nemotron-8B [5]. For commercial models, we select those with explicit grounding ability, including Gemini-2.5 (Pro/Flash) [9] and Seed-1.6-Vision-250815 [13].

4.2. Evaluation Settings

We employ a rigorous and standardized strategy for evaluation. For every data instance, we organize the input image and the description using a unified prompt template. This template accurately specifies critical details, including the reference viewpoint for spatial relationships, the allowed number of target objects to be output, and strict constraints on the output format. Unless otherwise specified in subsequent sections, all experiments are conducted using greedy decoding (set as temperature $= 0$). For the evaluation metric, we adopt the widely-used Accuracy@0.5, which represents the proportion of total samples where the Intersection over Union (IoU) between the ground-truth and predicted bounding box exceeds 0.5. The detailed prompt and results across various IoU thresholds are provided in the Appendix.

4.3. Evaluation Results

4.3.1. Main Results

Tab. 3 presents the evaluation results of all models on GroundingME. We explicitly disabling the thinking mode where supported. Our assessment yields three major observations. (1) Models demonstrate a significant performance gap on GroundingME, with the best model, Qwen3-VL-235B-A22B, achieving 45.1% accuracy, the majority score between 10% and 40%, and several with an accuracy only less than 10%, which strongly validates the challenge of our benchmark compared to existing ones. (2) Commercial models do not exhibit a pronounced advantage over open-source ones. We observe that even the best performed Seed-1.6-Vision-250815 closely follows the best open-source model at 42.6%, while Gemini-2.5 series show performance comparable only to mid-range open-source counterparts. (3) Model scale is a critical factor for performance. This scaling trend is consistently verified across model families, including Qwen3-VL-Dense (2B to 32B: 21.1% to 39.5%), Qwen3-VL-MoE (A3B to A22B: 35.7% to 45.1%), and Qwen2.5-VL 7B to 72B: 15.1% to 29.6%). This correlation underscores the importance of model size in achieving advanced visual grounding capabilities.

Table 3. **Evaluation results on GroundingME.** All models in this table are evaluated under the no-thinking mode setting if supported. All reported metrics in this table are Accuracy@0.5. The abbreviations for the subcategories are: App. (Appearance), Cmp. (Component), Txt. (Text), Sta. (State), Rel. (Relationship), Cnt. (Counting), Occ. (Occlusion), Sml. (Small), Avg. (Average). The best results are shown in **bold** and the second best is with underline. Detailed model specifications can be found in §4.1.

Model	Discriminative					Spatial			Limited			Rejection					Total
	App.	Cmp.	Txt.	Sta.	Avg.	Rel.	Cnt.	Avg.	Occ.	Sml.	Avg.	App.	Cmp.	Txt.	Sta.	Avg.	
Phi-4-Multimodal	3.8	0.0	0.0	0.0	1.0	0.7	0.7	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
Llama-4-Maverick	15.4	30.0	16.0	11.5	18.1	25.3	19.3	22.3	9.3	0.7	5.0	4.0	3.9	**12.0**	4.0	6.0	13.0
Llama-4-Scout	21.2	26.0	14.0	9.6	17.6	18.0	6.7	12.3	7.3	0.0	3.7	4.0	0.0	6.0	0.0	2.5	8.9
LLaVA-O.V.-1.5-8B	3.8	14.0	10.0	11.5	9.8	4.0	5.3	4.7	6.0	0.7	3.3	0.0	0.0	0.0	0.0	0.0	4.4
Mistral-3.2-24B	1.9	8.0	2.0	5.8	4.4	2.0	3.3	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7
Gemma-3-27B	3.8	0.0	0.0	1.9	1.5	0.7	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
Llama-Nemotron-8B	19.2	30.0	32.0	19.2	25.0	7.3	4.7	6.0	16.0	0.7	8.3	6.0	**5.9**	6.0	4.0	5.5	10.4
MiniCPM-V-4.5	7.7	12.0	6.0	5.8	7.8	4.7	3.3	4.0	8.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	4.0
InternVL3.5-8B	11.5	4.0	4.0	5.8	6.4	4.7	3.3	4.0	3.3	0.0	1.7	4.0	0.0	0.0	2.0	1.5	3.3
InternVL3.5-A28B	34.6	40.0	18.0	21.2	28.4	33.3	16.7	25.0	26.0	0.0	13.0	0.0	0.0	0.0	0.0	0.0	17.1
Keye-VL-1.5-8B	25.0	22.0	22.0	17.3	21.6	7.3	8.7	8.0	11.3	0.0	5.7	0.0	0.0	0.0	0.0	0.0	8.5
MiMo-VL-7B-RL	42.3	46.0	50.0	38.5	44.1	24.7	14.0	19.3	26.0	0.0	13.0	0.0	0.0	0.0	0.0	0.0	18.6
GLM-4.5V	50.0	58.0	58.0	46.2	52.9	54.7	29.3	42.0	48.0	10.7	29.3	0.0	0.0	0.0	2.0	0.5	32.1
Qwen2.5-VL-7B	23.1	36.0	46.0	23.1	31.9	12.7	16.0	14.3	28.0	0.7	14.3	0.0	0.0	2.0	0.0	0.5	15.1
Qwen2.5-VL-32B	34.6	58.0	66.0	32.7	47.5	48.7	31.3	40.0	35.3	0.0	17.7	0.0	0.0	0.0	0.0	0.0	26.9
Qwen2.5-VL-72B	50.0	50.0	66.0	28.8	48.5	52.0	28.7	40.3	46.7	0.7	23.7	4.0	3.9	4.0	0.0	3.0	29.6
Qwen3-VL-2B	44.2	36.0	66.0	32.7	44.6	11.3	12.0	11.7	21.3	36.0	28.7	0.0	0.0	0.0	0.0	0.0	21.1
Qwen3-VL-4B	55.8	58.0	74.0	38.5	56.4	34.7	22.0	28.3	45.3	48.7	47.0	0.0	0.0	0.0	0.0	0.0	33.9
Qwen3-VL-8B	55.8	68.0	80.0	42.3	61.3	32.7	20.0	26.3	56.0	16.0	36.0	0.0	0.0	0.0	0.0	0.0	31.0
Qwen3-VL-32B	**78.8**	**84.0**	82.0	55.8	**75.0**	61.3	33.3	47.3	60.7	7.3	34.0	0.0	0.0	0.0	0.0	0.0	39.5
Qwen3-VL-A3B	73.1	66.0	76.0	38.5	63.2	38.0	22.0	30.0	41.3	**52.0**	46.7	0.0	0.0	0.0	0.0	0.0	35.7
Qwen3-VL-A22B	71.2	74.0	**84.0**	50.0	69.6	62.7	36.7	49.7	**72.7**	35.3	**54.0**	0.0	0.0	0.0	0.0	0.0	**45.1**
Gemini-2.5-Pro	32.7	32.0	44.0	30.8	34.8	39.3	28.7	34.0	13.3	0.7	7.0	**10.0**	3.9	8.0	**6.0**	**7.0**	20.7
Gemini-2.5-Flash	40.4	40.0	34.0	30.8	36.3	28.7	21.3	25.0	22.7	3.3	13.0	0.0	0.0	0.0	0.0	0.0	18.7
Seed-1.6-V.-250815	59.6	80.0	36.0	**63.5**	59.8	**72.7**	**44.7**	**58.7**	70.0	15.3	42.7	0.0	0.0	2.0	2.0	1.0	42.6

4.3.2. Subtask Results

Analyzing the four distinct types of L-1 categories, we observe several key performance patterns. (1) Model performance exists in stratification across the different subtask types: models are generally most proficient at Discriminative category, followed by Spatial or Limited categories, while all models demonstrate poor performance on Rejection category. (2) Within the Discriminative category, a clear performance gain is tied to model scale, as seen in Qwen3-VL 2B to 32B: Average 44.6% to 75.0%), and performance on the State subcategory remains lower than the other three subcategories. (3) For the Spatial category, models are more proficient at the Relational subcategory, which requires qualitative positional ability, compared to the Counting subcategory, which requires quantitative assessment. Furthermore, performance in this category is more sensitive to model scale, with Qwen3-VL (2B to 32B) showing substantial improvement (Relationship: 11.3% to 61.3%; Average: 11.7% to 47.3%). More cross-dimensional analysis is provided in the Appendix.

5. Analysis

This section provides a detailed analysis of MLLM behaviors and capabilities on GroundingME. We first assess the performance gain achieved by enabling thinking mode (§5.1). Following this, we investigate a novel Test-Time Scaling strategy leveraging thinking trajectories (§5.2). Finally, we explore a data mixture training strategy designed to enhance the Rejection grounding capability (§5.3).

5.1. The Effectiveness of Thinking

The widely usage of thinking has demonstrated significant advantages in enhancing the complex reasoning capabilities of recent MLLMs. Therefore, we assess the impact of thinking on several representative MLLMs that natively supports thinking mode: Qwen3-VL (8B/32B/A3B/A22B), GLM-4.5V, MiMo-7B-RL-2508, and Seed-1.6-Vision-250815. Our experiments yield three main findings. (1) Thinking mode universally leads to better performance on GroundingME, with every tested model achieving a significant performance gain, ranging



Description
This object is a young child, likely a boy, with light-colored, possibly blonde or light brown, hair. **He is wearing a long-sleeved white shirt and black pants.** ...

Thinking Trajectory 1 (Correct)
Got it, let's look at the image. The description specifies a child with light hair, a long-sleeved white shirt, and black pants. ... **In the image, the child with light hair is wearing a white short-sleeve shirt and blue shorts. This does not match the description.** ... After checking all details, there is no child matching the required long-sleeve shirt and black pants. **Therefore, output null.**

Thinking Trajectory 2 (Wrong)
Got it, let's look at the image. The description is about a young child with light hair, white long-sleeve shirt, black pants. ... The child in the white shirt has light hair but is wearing blue shorts, not black pants. ... **Noting the discrepancy in sleeve length and pants color, but considering that the description might have an error.** ... Since this is the only child with light hair and a white shirt, **the bounding box is selected as [592, 307, 697, 542].**

Figure 4. **Case study of two different thinking trajectories of Qwen3-VL-235B-A22B-Thinking for the same description. The correct answer is to do rejection and the red bounding box shows the distractor.** The correct trajectory (Green) demonstrates rigorous adherence to the description, systematically identifying all attribute mismatches (e.g., short- vs. long-sleeve, blue vs. black pants) and correctly concluding with a null output. In contrast, the erroneous trajectory (Red) acknowledges the same discrepancies but compromises by speculating that the description may be in error, ultimately leading to an incorrect bounding box prediction.

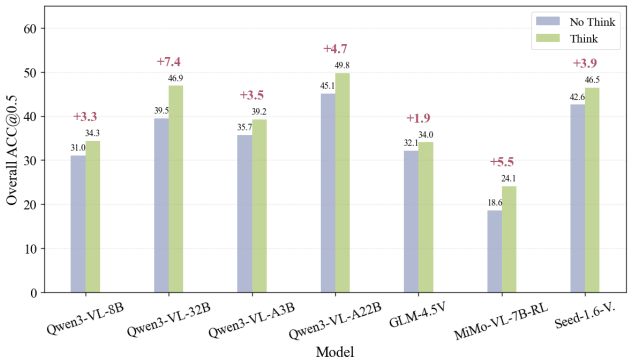


Figure 5. **Performance gain of different models by enabling thinking mode.** Subtask results are provided in the appendix.

Table 4. **Performance of our Test-time scaling method on GroundingME.** We compare the performance gain of Qwen3-VL-235B-A22B-Thinking when its 16 responses are selected by a judge model, with and without access to thinking trajectories.

Method	Judge Model	Total	Dis.	Spa.	Lim.	Rej.
Average	-	49.8	66.6	74.5	43.3	5.7
w/o CoT	Qwen3-VL-A3B	49.6	64.2	71.0	45.7	8.5
	Qwen3-VL-A22B	52.2	**69.1**	75.0	45.7	11.0
w/ CoT	MiMo-RL-0530	52.0	64.2	77.3	43.0	15.4
	Deepseek-R1	52.7	65.2	77.3	44.7	15.4
	Qwen3-VL-A22B	**54.3**	66.7	**79.7**	**46.3**	**15.9**

from 4.7% for GLM-4.5V to 7.4% for Qwen3-VL-32B. (2) Thinking Mode generally exhibits a negative effect on tasks relying more on perception, yet shows notable performance improvements for tasks prioritizing reasoning. Detailed subtask performance is provided in the Appendix. (3) Models can learn to reject by thinking. Under the no-thinking setting, most models exhibits 0% accuracy on the Rejection task, signifying a complete failure in executing any rejection behavior. Although performance on the rejection task still remains low under the thinking mode, all models successfully demonstrate some level of rejection behavior.

5.2. Test-Time Scaling by Thinking Quality

To gain a deeper understanding of how thinking contributes to model performance, we conduct a detailed case study focusing on the relationship between the quality of the generated thinking trajectory and the final grounding accuracy. As demonstrated in Fig. 4, different thinking trajectories for the same query can be highly divergent, leading to distinct outputs. We hypothesize that trajectories that are coherent, logically consistent, and strictly adhere to task instructions are more likely to lead to the correct answer.

To validate this hypothesis, we design a tailored Test-Time Scaling (TTS) method [7, 32, 41] with an LLM as the judge [18, 45–47]. For each sample, we use Qwen3-VL-235B-A22B-Thinking to generate 16 responses (temperature=0.7). We then employ a judge model to perform a "Best-of-16" selection: comparing the 16 responses in pairs, selecting the one with better thinking quality, and repeating until only one response remains. We test two text-only judges, DeepSeek-R1 [14] and MiMo-7B-RL-0530 [44], and one multimodal judge, Qwen3-VL-235B-A22B-Thinking. As baselines, we also test two multimodal judges, Qwen3-VL-235B-A22B-Thinking and Qwen3-VL-30B-A3B-Thinking, which only receive the final answers of responses without access to their thinking trajectories.

Based on the results presented in Tab. 4, we observe three major findings. (1) TTS based on thinking quality significantly boosts performance. Using Qwen3-VL-A22B as a judge with thinking trajectories improves performance across all subtasks, achieving a 4.5% total gain. (2) Even

Table 5. **In-domain performance of fine-tuned Qwen3-VL-8B-Instruct** on RefCOCOg validation split, our curated RefCOCOg_rej validation split, and the macro average of both datasets under different SFT data ratios (negative to positive).

Dataset	Origin	1:8	1:4	1:2	1:1	2:1
RefCOCOg_val	88.2	**90.4**	89.9	88.1	86.8	83.1
RefCOCOg_rej_val	30.5	83.5	87.9	92.3	94.8	**97.3**
Macro_Average	59.4	87.0	88.9	90.2	**90.8**	90.2

text-only judges prove effective. DeepSeek-R1 achieves a 2.9% performance gain, and MiMo-7B-RL-0530 yields a 2.2% gain. This demonstrates that good thinking trajectories alone without multimodal perception can improve accuracy. (3) Removing thinking trajectories degrades performance. When the multimodal judge Qwen3-VL-A22B is deprived of thinking trajectories and judges solely based on the image, description, and final answers, the TTS gain drops by 2.1%. Furthermore, when using Qwen3-VL-A3B as the judge, TTS does not bring any improvement.

5.3. Enhancing Rejection by Data Mixture

The performance of models on Rejection subtask exhibits a substantial performance disparity relative to their accuracy on positive samples. Motivated by prior work [2, 27] suggesting that compromised rejection capability stems from a scarcity of negative instances within the training corpus, we propose to investigate a data mixture training strategy for improving the model's rejection capability.

Considering the scale and accessibility, we utilize the RefCOCOg dataset as our base. We select 30,000 positive samples from its training split and construct 30,000 corresponding negative samples by modifying the description. We refer to this newly created set of negative samples as RefCOCOg_rej_train. These 60,000 instances serve as the source pool for generating various SFT datasets. To investigate the effect of different data mixture ratios, we randomly sample 30,000 instances from this pool to create five distinct fine-tuning datasets, with the negative-to-positive sample ratios set sequentially as 1:8, 1:4, 1:2, 1:1, and 2:1. Simultaneously, we construct 11,490 negative samples from the RefCOCOg validation split for evaluation, which we term RefCOCOg_rej_val. We then fine-tune Qwen3-VL-8B-Instruct for 3 epochs using these five mixture datasets.

We first evaluate the in-domain performance of the fine-tuned models on RefCOCOg_val (original positive samples) and RefCOCOg_rej_val (curated negative samples). From the results in Tab. 5, we find that: (1) Simple incorporation of negative data effectively enables the model to learn the rejection capability in visual grounding. (2) As the proportion of negative data increases in the fine-tuning set, the model's performance on the rejection task improves incrementally, while simultaneously impacting the original positive grounding performance to some extent. (3) The macro average accuracy across both negative and positive classes shows a significant performance gain of approximately 30% across various mixture ratios, powerfully confirming the efficacy of our data mixture training strategy.

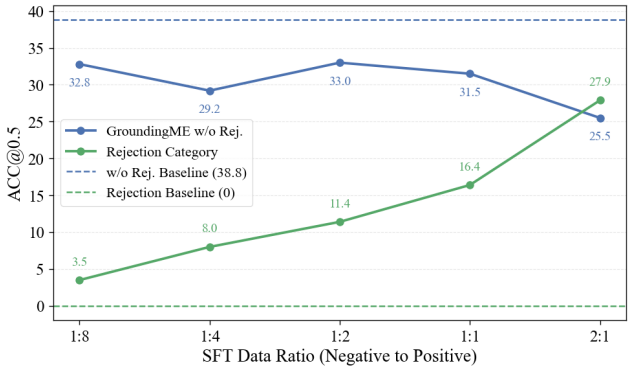


Figure 6. **Out-of-domain performance of fine-tuned Qwen3-VL-8B-Instruct** on GroundingME w/o Rejection and the Rejection category, as a function of SFT data ratio (negative to positive). Baseline means the performance before fine-tuning.

We further evaluate the out-of-domain performance of the fine-tuned models on GroundingME to explore the generalizability of the learned rejection capability. We denote the GroundingME benchmark excluding the rejection category as GroundingME w/o Rejection. As shown in Fig. 6, we observe that: (1) The performance trend on the Rejection category of GroundingME aligns with the in-domain results, showing a clear growth as the ratio of negative samples increases. (2) In contrast to the in-domain results, the performance on GroundingME w/o Rejection demonstrates a clear degradation compared to the pre-fine-tuned baseline from 38.8% to a max of 33.0%. This suggests that the rejection capability gained from simple data mixture does not generalize for free to higher-difficulty, out-of-domain scenarios, pointing towards a critical challenge that necessitates further investigation in future work.

6. Conclusion

In this work, we introduce GroundingME, a challenging benchmark that rigorously evaluates MLLMs' visual grounding capabilities through carefully curated samples across multiple dimensions. Our evaluation reveals that current MLLMs, despite strong performance on existing benchmarks, still struggle with more challenging scenarios. We further show that strategies including test-time scaling based on thinking quality and rejection-focused data-mixture training offer promising directions for improvement. We hope GroundingME will drive progress toward more capable and trustworthy visual grounding systems.

Acknowledgments

This paper is supported by NSFC project 62476009 and the Open Project Fund of the State Key Laboratory of Multimedia Information Processing (Project No. SKLMIP-KF-2025-01). We thank the Xiaomi MiMo team for their helpful discussions and support. We also thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *ArXiv preprint*, abs/2503.01743, 2025. 5

[2] Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29612–29622, 2025. 8

[3] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *ArXiv preprint*, abs/2509.23661, 2025. 5

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *ArXiv preprint*, abs/2502.13923, 2025. 1, 2, 5

[5] Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. *ArXiv preprint*, abs/2505.00949, 2025. 5

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18392–18402, 2023. 2

[7] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher R'e, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *ArXiv preprint*, abs/2407.21787, 2024. 2, 7

[8] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 513–524, 2025. 2, 3, 4, 5

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv preprint*, abs/2507.06261, 2025. 2, 4, 5

[10] Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. Ref-adv: Exploring mllm visual reasoning in referring expression tasks. In *The Fourteenth International Conference on Learning Representations*. 2, 3

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024. 5

[12] Gemini Team. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805, 2023. 1

[13] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *ArXiv preprint*, abs/2505.07062, 2025. 2, 5

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 7

[15] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4044–4054, 2023. 2, 3

[16] 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)*, pages 787–798, 2014. 2, 3

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003, 2023. 3

[18] Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. Vl-rewardbench: a challenging benchmark for vision-language generative reward models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24657–24668, 2025. 7

[19] Xiangtai Li, Tao Zhang, Yanwei Li, Haobo Yuan, Shihao Chen, Yikang Zhou, Jiahao Meng, Yueyi Sun, Shilin Xu, Lu Qi, et al. Denseworld-1m: Towards detailed dense grounded caption in the real world. *ArXiv preprint*, abs/2506.24102, 2025. 3

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1

[21] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *IEEE Conference on Computer Vision and Pat-*

tern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 4185–4194, 2019. 2, 3

[22] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3

[23] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20, 2016. 2, 3

[24] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016. 2, 3

[25] OpenAI. Gpt-4v(ision) system card, 2023. 1

[26] Georgios Pantazopoulos and Eda B Özyiğit. Towards understanding visual grounding in visual language models. *ArXiv preprint*, abs/2509.10345, 2025. 1, 3

[27] Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know" no"better: A data-driven approach for enhancing negation awareness in clip. *ArXiv preprint*, abs/2501.10913, 2025. 8

[28] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020. 2, 3

[29] Heqian Qiu, Hongliang Li, Taijin Zhao, Lanxiao Wang, Qingbo Wu, and Fanman Meng. Refcrowd: Grounding the target in crowd with referring expressions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4435–4444, 2022. 2, 3

[30] Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. Aligning and prompting everything all at once for universal visual perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13193–13203, 2024. 3

[31] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *ArXiv preprint*, abs/2406.14555, 2024. 2

[32] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *ArXiv preprint*, abs/2408.03314, 2024. 2, 7

[33] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *ArXiv preprint*, abs/2503.19786, 2025. 5

[34] GLM-V Team. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. *ArXiv preprint*, abs/2507.01006, 2025. 2, 5

[35] Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl technical report. *ArXiv preprint*, abs/2507.01949, 2025. 5

[36] Mistral Team. Mistral 7b, 2023. 5

[37] Jiaqi Wang, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Bao Ge, and Shu Zhang. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence*, 2024. 2

[38] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7907–7915, 2025. 3

[39] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *ArXiv preprint*, abs/2508.18265, 2025. 2, 5

[40] 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 *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 2, 3

[41] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. 2024. 2, 7

[42] Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 3

[43] LLM-Core Xiaomi. Mimo-vl technical report. *ArXiv preprint*, abs/2506.03569, 2025. 2, 5

[44] LLM-Core Xiaomi. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *ArXiv preprint*, abs/2505.07608, 2025. 7

[45] Tianyi Xiong, Yi Ge, Ming Li, Zuolong Zhang, Pranav Kulkarni, Kaishen Wang, Qi He, Zeying Zhu, Chenxi Liu, Ruibo Chen, et al. Multi-crit: Benchmarking multimodal judges on pluralistic criteria-following. *arXiv preprint arXiv:2511.21662*, 2025. 7

[46] Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13618–13628, 2025.

[47] Tianyi Xiong, Shihao Wang, Guilin Liu, Yi Dong, Ming Li, Heng Huang, Jan Kautz, and Zhiding Yu. Phycritic: Multimodal critic models for physical ai. *arXiv preprint arXiv:2602.11124*, 2026. 7

[48] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016. 2, 3

[49] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *ArXiv preprint*, abs/2509.18154, 2025. 5

[50] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. Large language models for robotics: A survey. *ArXiv preprint*, abs/2311.07226, 2023. 2

[51] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 3