

Revisiting Referring Expression Comprehension Evaluation in the Era of Large Multimodal Models

Jierun Chen^{1,*} Fangyun Wei^{2,*†} Jinjing Zhao³ Sizhe Song¹
Bohuai Wu¹ Zhuoxuan Peng¹ S.-H. Gary Chan¹ Hongyang Zhang⁴

¹ HKUST ² Microsoft Research Asia ³ The University of Sydney ⁴ University of Waterloo

Abstract

Referring expression comprehension (REC) involves localizing a target instance based on a textual description. Recent advancements in REC have been driven by large multimodal models (LMMs) like CogVLM, which achieved 92.44% accuracy on RefCOCO. However, this study questions whether existing benchmarks such as RefCOCO, RefCOCO+, and RefCOCOg, capture LMMs' comprehensive capabilities. We begin with a manual examination of these benchmarks, revealing high labeling error rates: 14% in RefCOCO, 24% in RefCOCO+, and 5% in RefCOCOg, which undermines the authenticity of evaluations. We address this by excluding problematic instances and reevaluating several LMMs capable of handling the REC task, showing significant accuracy improvements, thus highlighting the impact of benchmark noise. In response, we introduce Ref-L4, a comprehensive REC benchmark, specifically designed to evaluate modern REC models. Ref-L4 is distinguished by four key features: 1) a substantial sample size with 45,341 annotations; 2) a diverse range of object categories with 365 distinct types and varying instance sizes from 30 to 3,767; 3) lengthy referring expressions averaging 24.2 words; and 4) an extensive vocabulary comprising 22,813 unique words. We evaluate a total of 24 large models on Ref-L4 and provide valuable insights. The cleaned versions of RefCOCO, RefCOCO+, and RefCOCOg, as well as our Ref-L4 benchmark and evaluation code, are available at <https://github.com/JierunChen/Ref-L4>.

1. Introduction

Referring expression comprehension (REC) [37, 66] involves the task of localizing a specific target instance based on a given textual description. The advancement of REC has been significantly propelled by the superior language processing capabilities of large language models

*Equal contribution.

†Corresponding author.



Figure 1. Examples from the previous RefCOCO, RefCOCO+, and RefCOCOg datasets, and our Ref-L4 benchmark. Current datasets typically contain overly brief expressions and simple scenes with large target instance. In contrast, our Ref-L4 benchmark features more detailed expression, a broader variety of scenes with more categories, and instances across various scales, particularly the challenging small instances. We also rigorously identity labeling errors within current datasets, with an example shown in the upper-right subfigure.

(LLMs) [36, 56, 57]. This progress is particularly evident in the exceptional performance of large multimodal models (LMMs) [2, 9, 12, 16, 58, 82] on well-known benchmarks such as RefCOCO [73], RefCOCO+ [73], and RefCOCOg [35]. These models have demonstrated remarkable accuracy, with CogVLM [63], for instance, achieving an impressive accuracy rate of 92.44% on the RefCOCO benchmark.

This paper begins with a critical question: do existing

Table 1. Statistics of the labeling error rates for RefCOCO, RefCOCO+, and RefCOCOg, respectively. For each benchmark, the statistics are conducted on the combination of the validation and test sets.

Benchmark	Annotations	Errors	Labeling Error Rate
RefCOCO	21,586	3,054	14%
RefCOCO+	21,373	5,201	24%
RefCOCOg	14,498	675	5%

Table 2. Performance of four LMMs on REC task across our cleaned and the original versions of RefCOCO, RefCOCO+, and RefCOCOg benchmarks, using the conventional accuracy as the evaluation metric. The evaluation is performed on the combination of the validation and test sets for each benchmark. †: models fine-tuned on the specific dataset.

Benchmark	ONE-PEACE†	OFA-L†	OFA-L	Qwen-VL	CogVLM-Grounding
RefCOCO	92.15	89.85	85.13	88.51	92.44
RefCOCO (Cleaned)	94.11 (+1.96)	92.06 (+2.22)	87.95 (+2.81)	90.68 (+2.18)	94.58 (+2.13)
RefCOCO+	88.14	85.06	77.56	82.52	88.55
RefCOCO+ (Cleaned)	90.79 (+2.66)	87.38 (+2.32)	80.50 (+2.94)	85.60 (+3.08)	91.43 (+2.87)
RefCOCOg	89.18	84.77	79.25	85.11	90.67
RefCOCOg (Cleaned)	90.75 (+1.57)	86.39 (+1.62)	80.89 (+1.64)	86.79 (+1.68)	92.36 (+1.68)

Table 3. Comparison between our Ref-L4 benchmark and other REC benchmarks, including RefCOCO, RefCOCO+, and RefCOCOg. For the latter three benchmarks, we combine their validation and test sets for statistics. The instance size and image size are represented by their respective square roots. Avg. length: average length of annotations. Vocab.: vocabulary size.

Benchmark	Images	Instances	Annotations	Categories	Avg. Length	Instance Size	Normalized Inst. Size	Image Size	Vocab. Size
RefCOCO	3,000	7,596	21,586	71	3.6	105 - 607	0.17 - 1.0	230 - 640	3,525
RefCOCO+	3,000	7,578	21,373	71	3.6	105 - 607	0.17 - 1.0	230 - 640	4,387
RefCOCOg	3,900	7,596	14,498	78	8.4	83 - 610	0.22 - 1.0	277 - 640	5,050
Ref-L4 (Ours)	9,735	18,653	45,341	365	24.2	30 - 3,767	0.05 - 1.0	230 - 6,606	22,813

REC benchmarks truly capture the comprehensive capabilities of LMMs? The foundational benchmarks, RefCOCO, RefCOCO+, and RefCOCOg, were introduced sequentially in 2015, 2016, and 2016, respectively. In RefCOCO, the referring expressions are notably succinct, ranging from single words like “yellow” (see Fig. 1 (a)) and “lady” to brief descriptions such as “far left person” and “white shirt”. RefCOCO+ intentionally excludes locational prepositions commonly found in RefCOCO, favoring short yet semantically rich expressions like “man in hat” (see Fig. 1 (a)) and “plastic cup with just ice”. RefCOCOg provides a bit more elaborate annotations, including examples such as “a bench that is close to the mopads” and “the dog with a paw on the windowsill”. These brief annotations align the REC task more closely with the straightforward mapping from textual labels to a visual element. This similarity raises concerns about the adequacy of these benchmarks in thoroughly evaluating the nuanced language comprehension and reasoning abilities of modern LMMs.

Regarding visual complexity, the RefCOCO, RefCOCO+, and RefCOCOg benchmarks represent a restricted set of object categories (fewer than 80), limiting their reflection of real-world diversity and their ability to assess the extensive world knowledge acquired by modern LMMs. Additionally, the target instances in these datasets are often large relative to the total image (see Fig. 1 (a)), simplifying the task compared to more challenging scenarios involving complex or wide-angle scenes with small objects. Lastly, a significant limitation of these datasets is the prevalence of labeling errors, with an example shown in the upper-right subfigure in Fig. 1 (a).

Labeling Error Rates of Existing Benchmarks. To be-

gin, we manually assess the labeling error rates of the validation and test sets in RefCOCO, RefCOCO+, and RefCOCOg, discovering a high error rate across these benchmarks. The labeling errors include, typos, misalignment between referring expressions and target instances, as well as inaccurate bounding box annotations, as depicted in Sec. 6. As illustrated in Tab. 1, the labeling error rates for RefCOCO, RefCOCO+, and RefCOCOg are 14%, 24%, and 5%, respectively, indicating that evaluations performed on these benchmarks may lack authenticity.

Reevaluation on RefCOCO, RefCOCO+ and RefCOCOg. In response, we manually exclude the problematic instances from the validation and test sets of RefCOCO, RefCOCO+, and RefCOCOg. Subsequently, we reevaluate four LMMs capable of handling the REC task—namely ONE-PEACE [61], OFA-L [60], Qwen-VL [1], and CogVLM-Grounding [63]—on both the cleaned and original versions of these datasets, as shown in Tab. 2. Across all models and cleaned benchmarks, we observe a significant accuracy improvement, ranging from 1.57 to 3.08, compared to their performance on the original versions. This demonstrates that noise in the benchmarks has affected the valid evaluation and comparisons the models’ true capabilities. *To support further research in the REC field, we will release the cleaned versions of RefCOCO, RefCOCO+, and RefCOCOg.* Moreover, the accuracy gains after filtering the noise highlight the performance plateau on current benchmarks, motivating us to propose a new, more challenging benchmark to better assess and distinguish REC models.

Ref-L4: A Comprehensive REC Benchmark for Modern LMM Evaluation. We present Ref-L4, where L4 signifies four key aspects: a Large number of testing

samples, Large diversity in object categories and instance scales, Long referring expressions, and a Large vocabulary. These features make Ref-L4 a comprehensive benchmark for assessing the REC capabilities of contemporary LMMs. Tab. 3 provides a detailed comparison between Ref-L4 and other benchmarks including RefCOCO, RefCOCO+, and RefCOCOg. Our Ref-L4 benchmark stands out due to the following characteristics:

- *Large-Scale.* Ref-L4 includes 9,735 images, 18,653 unique instances, and a total of 45,341 annotations, significantly surpassing RefCOCO, RefCOCO+, and RefCOCOg. For instance, RefCOCOg offers 3,900 images, 7,596 instances, and 14,498 annotations.
- *High Diversity.* Ref-L4 features 365 unique categories. Since the RefCOCO series derive from the COCO 2014 dataset, they encompass up to 78 categories. Additionally, our benchmark covers a wider range of instance sizes, from 30 to 3,767, measured by the square root of the instance area. Regarding normalized instance size, our Ref-L4 incorporates more challenging small instances, reaching down to 0.05, calculated as the square root of the instance area relative to the total image area.
- *Lengthy Referring Expressions.* Each referring expression in Ref-L4 is a detailed description of a specific instance, with lengths ranging from 3 to 117 words and an average of 24.2 words. In comparison, the average annotation lengths in RefCOCO, RefCOCO+, and RefCOCOg are 3.6, 3.6, and 8.4 words, respectively. Examples can be found in Fig. 1 (b).
- *Extensive Vocabulary.* Due to the detailed nature of the referring expressions, Ref-L4 boasts a large vocabulary of 22,813 words, which is four to six times larger than those of RefCOCO, RefCOCO+, and RefCOCOg.

Evaluation on Ref-L4. We conduct an evaluation of 24 representative LMMs that can perform the REC task. In addition to the standard accuracy metric, which considers predictions with an IoU greater than 0.5 as accurate ($\text{Acc}_{0.5}$), we also report accuracies at higher IoU thresholds: $\text{Acc}_{0.75}$ and $\text{Acc}_{0.9}$. Furthermore, we introduce a mean accuracy (mAcc), calculated as the average accuracy from $\text{Acc}_{0.5}$ to $\text{Acc}_{0.95}$ in increments of 0.05. To gain deeper insights into the models’ capabilities, we conduct a detailed analysis of REC performance across different instance scales and categories. *The Ref-L4 benchmark and the evaluation code will be made available to the community.*

2. Related Work

REC and Its Benchmarks. Referring Expression Comprehension (REC) [17, 20, 37, 42, 46, 77, 83] is a task that involves identifying a specific object within an image based on a given referring expression. Unlike object detection [4, 22, 28, 50, 53], which operates within fixed categories and a single visual modality, REC necessitates

understanding free-form text to locate objects of any category. Phrase Grounding [15, 26, 33, 43, 62, 68, 78] is similar but typically involves shorter phrases and identifies multiple regions, whereas REC requires parsing longer expressions to pinpoint a single unique region. This complexity makes REC an ideal task for evaluating emerging large multimodal models. Referring Expression Segmentation (RES) [25, 30, 68] requires pixel-level prediction, but its large-scale labeling is too costly compared to the REC benchmarks. Current REC benchmarks such as RefCOCO [73], RefCOCO+[73], and RefCOCOg[35] include tens of thousands of annotations but are limited by their short expression lengths—averaging 3.6, 3.6, and 8.4 words, respectively. Additionally, they encompass fewer than 80 categories, lacking real-world diversity. Other REC benchmarks [3, 7, 8, 10, 11, 13, 18, 23, 32, 47, 52, 59, 65, 69] are often designed for specific scenarios. For example, CLEVR-Ref+[32] focuses on simple objects like boxes, spheres, and cylinders. SK-VG[8] integrates prior scene knowledge as additional input, while RefCrowd [47] targets identifying a person within a crowd. By contrast, we introduce Ref-L4, a more general and comprehensive benchmark encompassing 365 categories and 45,341 annotations. Ref-L4 features expressions averaging 24.2 words and a vocabulary of 22,813 words, facilitating the accurate evaluation of REC models on complex expressions and diverse objects.

REC Models. The evolution of REC models has transitioned from specialized models [19, 31, 55, 70, 74, 84, 86] to generalist models or large multimodal models (LMMs)[1, 5, 14, 21, 29, 34, 44, 45, 54, 61, 63, 64, 67, 75, 76, 79, 80]. Notable examples of these LMMs include CogVLM-Grounding[63], SPHINX [14, 29], ONEPEACE [61], Qwen-VL-Chat [1], MiniGPTv2 [5], and Lenna [67]. These models, benefiting from larger model sizes and extensive training on diverse datasets, exhibit remarkable performance on conventional REC datasets. For example, CogVLM-Grounding achieves an accuracy of 94.58% on RefCOCO (cleaned). Additionally, the performance gap among models is shrinking, with many LMMs surpassing 90% accuracy. This performance saturation raises concerns about the adequacy of current REC benchmarks for making meaningful comparisons. In response, we propose Ref-L4, a more comprehensive and challenging benchmark. We have also conducted rigorous evaluations of 24 LMM models, offering holistic comparisons that highlight their weaknesses and suggest directions for improvement.

3. Ref-L4

3.1. Benchmark Creation

Data Sources. Our benchmark is derived from two sources: 1) our cleaned validation and test sets of the RefCOCO [73],

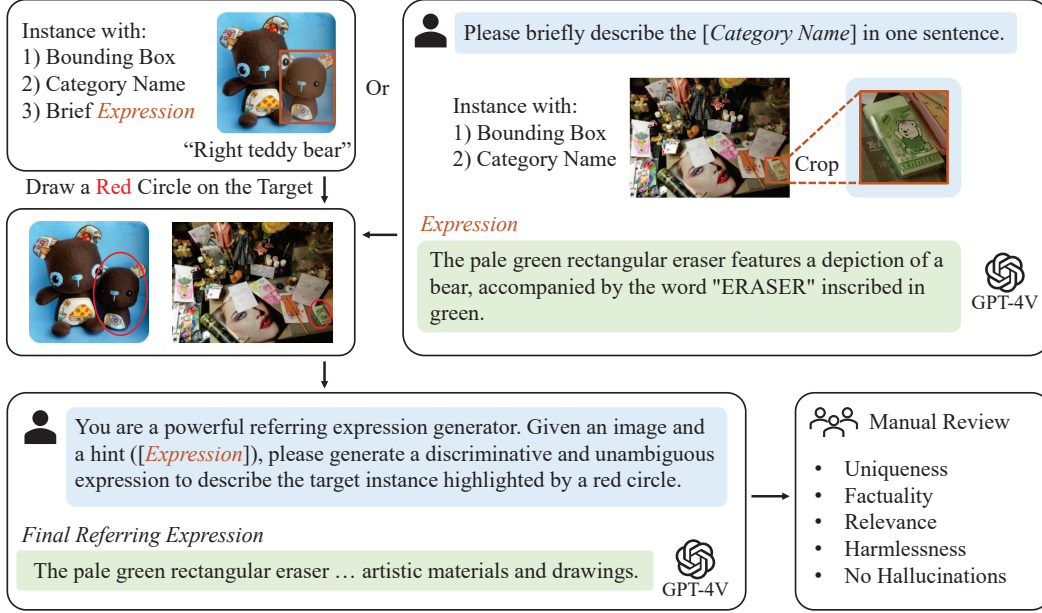


Figure 2. Pipeline of generating a referring expression for a target instance.

RefCOCO+ [73], and RefCOCOg [35] datasets; and 2) the test set from the large-scale object detection dataset Objects365 [53]. The Objects365 dataset provides a broader range of categories, varying instance sizes, higher image resolutions, and more intricate scenes. In the RefCOCO series, each instance includes a bounding box, a category name, and an extremely brief expression like “right teddy bear”. In contrast, the Objects365 benchmark labels each instance with mainly a bounding box and the relevant category.

For the RefCOCO (cleaned) series, we begin by consolidating duplicate images and instances, resulting in a subset of 6,502 images containing 14,186 unique instances. For Objects365, we select samples from its testing set based on several criteria: 1) each image has both height and width greater than 800 pixels; 2) each image is sufficiently complex, containing more than 10 categories and 20 instances; 3) each instance has a square normalized size $\sqrt{(hw)/(HW)}$ greater than 0.05, where (h, w) represents the instance size and (H, W) denotes the image size; 4) we randomly sample N instances for each of the 365 classes defined in Objects365, with $N = \min(35, \text{the number of instances for the specific class})$; 5) we review and exclude instances with erroneous bounding box annotations or those difficult to describe uniquely. For a few rare classes, we relax criterion-1 to 512 pixels and criterion-2 to 10 instances. Consequently, we collect 3,233 images and 4,467 instances from Objects365. Overall, our Ref-L4 benchmark comprises 9,735 images and 18,653 instances, sourced from the RefCOCO series and Objects365.

Referring Expression Generation. Given a target in-

stance and its corresponding image, we leverage GPT-4V with human reviewers in the loop to generate its precise and detailed referring expressions. Fig. 2 illustrates the three-step generation process:

Step-1: Each instance in the Objects365 dataset is linked to a bounding box and a *category name*. We begin by cropping these instances from the original images. Next, we input each cropped area along with the prompt detailed in Sec. 8.1 into GPT-4V to produce a context-independent description. For instances from the RefCOCO series, this step is omitted as each instance already has a brief expression.

Step-2: Drawing inspiration from recent studies on GPT-4V [71], where GPT-4V is able to pay more attention to instances highlighted by a red circle within an image, we similarly encircle the target instance in red to facilitate GPT-4V in generating a context-aware referring expression. Following this, as depicted in Fig. 2, we process the image and use the prompt outlined in Sec. 8.2 to generate a context-aware referring expression for each instance. We instruct GPT-4V to describe various features such as color, size, position, and context. Additionally, we provide a hint (the context-independent description from Step-1) in the prompt to mitigate hallucination issues, resulting in more accurate descriptions.

Step-3: We manually review all generated referring expressions to correct any hallucination issues. We ensure that each expression uniquely describes the instance and is factual, accurate, and harmless.

Annotation Expansion. To date, we have compiled 18,653 unique referring expressions, each describing a distinct instance. To assess the robustness of REC models to

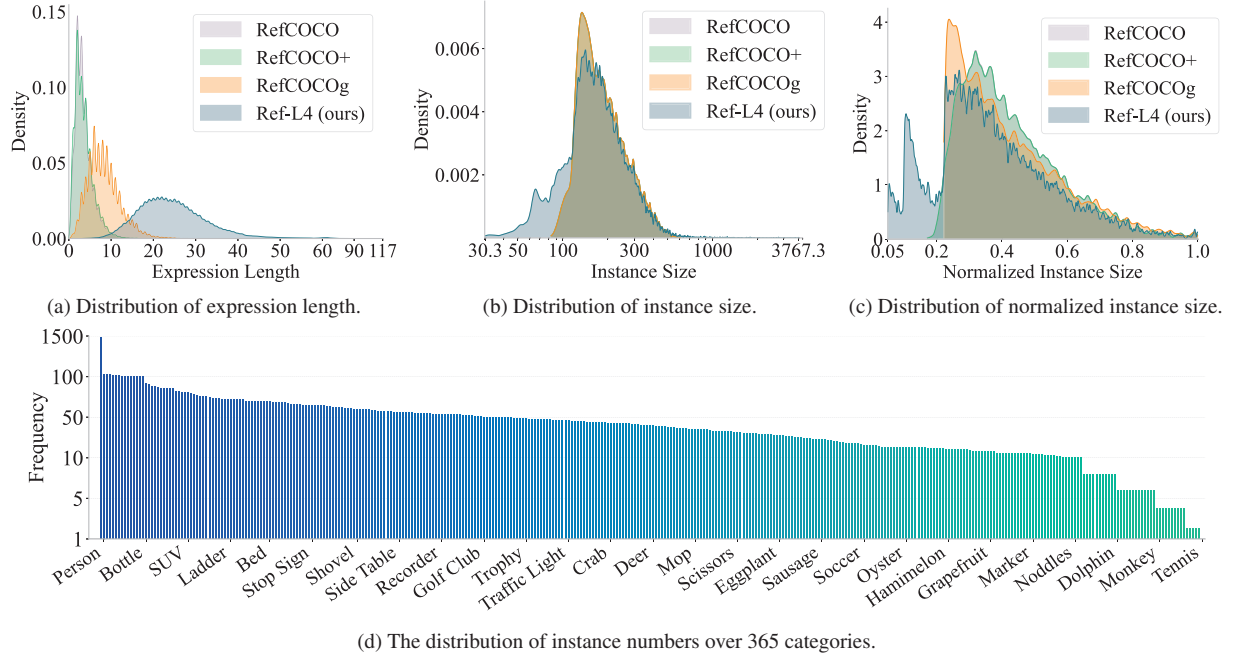


Figure 3. Analysis of referring expression length, (normalized) instance size, and category distribution.

diverse language inputs, we employ a two-stage rephrasing process to expand our benchmark: 1) utilizing GPT-4 with the prompt detailed in Sec. 8.3, to generate rephrased versions of each expression; 2) conducting a manual review to ensure that the rephrased expressions are unique, factual, relevant, and harmless.

- **Uniqueness:** Discard pairs if expressions could refer to more than one region without additional context.
- **Factualness:** Confirm that any mentioned attributes (color, shape, size, texture, position, relation, *etc.*) align with what is visible in the image.
- **Relevance:** Discard unnecessary details irrelevant to identifying the region.
- **Harmlessness:** Ensure that the expressions are neutral and do not contain inappropriate language that might imply stereotypes, violence, offense, or other potentially harmful perspectives.

Following the above guidelines, we ensure each expression reviewed independently by two reviewers for consistency. Only samples with consensus from both reviewers are retained. As a result, our final Ref-L4 benchmark encompasses 9,735 images with 45,341 referring expressions, each accurately describing one of the 18,653 unique instances. Examples can be found in Fig. 1 (b) and Sec. 7.

3.2. Analysis

Expression Length. Fig. 3a illustrates the distribution of expression lengths across four different datasets: RefCOCO, RefCOCO+, RefCOCOg, and our Ref-L4. Due

to the high overlap of data samples, RefCOCO and RefCOCO+ exhibit similar distributions, with a high density of shorter expressions peaking at around 3.6 words. RefCOCOg features slightly longer expressions on average, peaking at approximately 8.4 words. In contrast, our Ref-L4 displays a significantly different distribution, with expressions ranging much longer, peaking at around 24.2 words and having a long tail extending up to 117 words. This suggests that our Ref-L4 benchmark is structured to challenge current REC models, requiring them to comprehend and reason across more intricate and detailed descriptions. Such lengthier expressions also helps to uniquely referring to small objects within complex or wide-angle scenes, thereby reducing ambiguity.

Instance Size. We present the density plots comparing the instance sizes across four benchmarks. We define the instance size as the square root of the bounding box area \sqrt{hw} (see Fig. 3b), as well as a normalized version $\sqrt{(hw)/(HW)}$ (see Fig. 3c), where (h, w) represents the dimensions of the instance and (H, W) represents the dimensions of the image. While current benchmarks include only medium to large instances, our Ref-L4 benchmark presents a broader distribution range, with a particular emphasis on more challenging small instances.

Categories. Our Ref-L4 benchmark comprises 18,653 instances spanning 365 distinct categories, providing more complex and diverse evaluation scenarios. In contrast, RefCOCO and RefCOCO+ consists of 71 categories, while RefCOCOg covers 78 categories. Fig. 3d presents the dis-

Table 4. Scene diversity across 20 consolidated categories, predicted by GPT-4o and manually corrected, based on the combined validation and test sets.

Category	Percentage(%)	Category	Percentage(%)
Residential & Domestic Spaces	19.68	Entertainment	2.88
Catering & Dining	16.36	Recreational Facilities	2.46
Urban Scenes & Streetscapes	9.14	Water & Maritime Scenes	2.43
Transportation & Transit	8.89	Industrial & Workplaces	1.92
Sports & Exercise	8.71	Outdoor & Adventure	1.75
Wildlife	6.25	Hospitality, Resorts & Lodging	1.28
Commercial & Retail Spaces	5.18	Infrastructure & Public Services	1.03
Educational & Cultural Facilities	4.42	Health & Care Facilities	0.51
Agriculture & Rural	3.79	Natural Landscapes	0.11
Parks & Outdoor Leisure	3.16	Scientific Interest	0.05

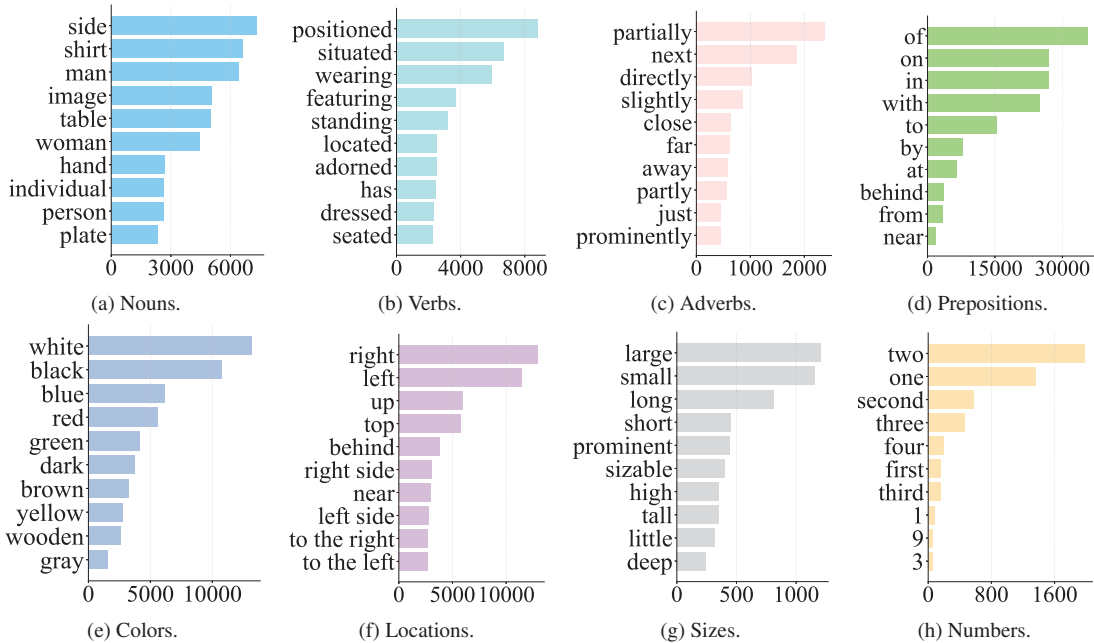


Figure 4. The frequency of the 10 most frequently used words in each part-of-speech category, as parsed using the SpaCy library.

tribution of instances among these 365 categories. Notably, the ten categories with the highest number of instances are “Person”, “Chair”, “Hat”, “Desk”, “Lamp”, “Cabinet/shelf”, “Car”, “Sneakers”, “Handbag/Satchel”, and “Flag”.

Scenes. We provide a detailed scene analysis on our benchmark. We start by referencing the 365 scene categories from the Places365 benchmark [85], known for being the most extensive dataset in scene recognition. These 365 categories are then consolidated into 20 broader groups using GPT-4o. Each image in our benchmark is processed by GPT-4o to predict its corresponding scene category, with manual corrections applied to ensure accuracy. The resulting statistics on scene diversity are summarized in the Tab. 4, with the combined validation and test sets used for this analysis.

Vocabulary. Our benchmark’s referring expressions

comprise a vocabulary totaling 22,813 unique words. This is significantly larger than the vocabulary sizes of RefCOCO, RefCOCO+, and RefCOCOg, which are 3,525, 4,387, and 5,050 words, respectively. Fig. 4 illustrates the 10 most frequently used nouns, verbs, adverbs, and prepositions, along with nouns indicating colors, locations, sizes, and numbers across all annotations.

3.3. Evaluation

Evaluation Metrics. We propose three distinct evaluation protocols:

1. *Accuracy.* This is the conventional metric used in REC. For a given referring expression and corresponding image, the target instance is considered successfully localized if the IoU between the predicted bounding box and the ground truth exceeds 0.5. Accuracy is then calculated as the ratio of successfully localized samples to the

Table 5. Performance evaluation across 24 models on our Ref-L4 benchmark. NVIDIA A100 GPUs (80G) are utilized. The symbol † denotes models that outputs segmentation masks.

Model	Val+Test				Val	Test	Small Size		Medium Size		Large Size	
	Acc _{0.5}	Acc _{0.75}	Acc _{0.9}	mAcc	mAcc	mAcc	Acc _{0.5}	mAcc	Acc _{0.5}	mAcc	Acc _{0.5}	mAcc
GPT-4V [38–40]	9.91	1.19	0.12	2.88	2.96	2.85	2.13	0.49	10.29	2.78	14.93	4.83
KOSMOS-2 [41]	48.53	38.34	17.54	34.72	34.89	34.64	24.19	11.63	46.95	32.91	69.32	54.98
OFA-Tiny [60]	55.21	43.22	27.70	41.44	41.53	41.40	17.91	11.49	65.13	49.00	64.46	49.61
OFA-Large [60]	72.53	62.31	45.02	59.17	59.42	59.07	40.13	27.07	81.03	66.49	80.78	69.36
Ferret-7b [72]	57.54	42.44	21.01	40.29	40.31	40.28	30.93	14.57	62.40	43.72	68.18	52.92
Ferret-13b [72]	64.44	49.04	27.46	46.88	47.31	46.71	36.46	17.88	70.50	51.86	73.92	59.09
GroundingGPT [27]	60.84	40.48	12.00	38.19	38.42	38.09	24.43	10.28	67.67	41.04	75.09	53.47
Shikra-7b [6]	65.06	39.62	10.45	38.60	38.91	38.47	43.91	18.50	75.98	46.27	60.60	39.34
Lenna [67]	65.90	58.55	45.58	55.69	55.88	55.60	31.02	23.48	72.90	61.53	78.72	68.66
MiniGPTv2 [5]	66.93	50.50	25.30	47.15	47.43	47.03	32.99	14.85	73.67	51.16	79.52	63.53
Qwen-VL-Chat [1]	73.80	58.05	37.16	55.94	56.18	55.83	47.66	26.26	79.80	61.06	82.01	68.37
ONE-PEACE [61]	70.82	60.09	36.12	55.07	55.49	54.89	22.18	13.98	83.26	63.39	83.81	70.04
SPHINX-MoE [14]	66.23	44.90	15.32	42.38	42.80	42.21	39.48	16.39	72.97	46.38	73.55	54.17
SPHINX-MoE-1k [14]	74.45	62.70	38.85	58.07	58.35	57.95	58.96	37.61	77.80	61.53	79.70	66.77
SPHINX [29]	74.78	53.65	21.15	50.09	50.33	49.99	48.82	22.08	80.56	54.10	83.27	63.34
SPHINX-1k [29]	78.52	62.17	32.95	57.57	57.91	57.42	59.48	33.21	82.95	61.82	84.40	67.68
SPHINX-v2-1k [29]	81.31	70.49	46.59	65.39	65.67	65.27	65.23	43.43	84.00	68.45	88.21	75.91
CogVLM-Grounding [63]	81.70	70.77	48.35	66.09	66.25	66.02	75.06	52.85	86.43	71.31	77.91	66.25
PixelLM-7B† [51]	41.83	27.57	13.32	27.10	27.09	27.11	8.25	4.05	43.90	27.33	62.72	43.64
PixelLM-13B† [51]	49.89	35.37	18.42	34.10	34.52	33.92	17.05	8.54	53.40	35.48	67.59	50.34
LISA-Explanatory† [24]	65.12	52.35	38.26	50.77	50.89	50.72	39.11	27.16	70.03	54.61	75.25	61.09
LISA† [24]	66.23	54.02	39.73	52.18	52.44	52.07	39.24	27.49	71.17	56.05	77.01	63.22
PSALM† [81]	67.26	58.22	44.11	55.46	55.68	55.37	37.35	28.43	75.06	61.79	74.97	63.74
GlaMM† [48]	71.90	60.27	45.15	57.89	58.16	57.78	47.07	34.36	77.17	62.28	80.50	67.14

total number of samples, referred to as $\text{Acc}_{0.5}$ in this work. To better assess the localization capabilities of modern REC models, we also report accuracies at higher IoU thresholds: $\text{Acc}_{0.75}$, $\text{Acc}_{0.9}$, and mAcc , which is the average accuracy from $\text{Acc}_{0.5}$ to $\text{Acc}_{0.95}$ in increments of 0.05.

2. *Scale-Aware Performance.* To gain deeper insights into model capabilities, we report performance based on instance sizes: small, medium, and large. The size of an instance is defined as the square root of its area, $\sqrt{(hw)}$, where (h, w) are the dimensions of the instance. Small instances are those with a size less than 128, medium instances are between 128 and 256, and large instances exceed 256. In total, there are 9345, 23280, and 12716 referring expressions describing 2,954 small, 10,442 medium, and 5,257 large instances, respectively.
3. *Per-Category Performance.* Our benchmark encompasses a wide range of categories, up to 365 in total. We provide an evaluation protocol to assess performance on a per-category basis.

Benchmark Division. Modern large multimodal models (LMMs) that are able to handle the REC task typically use unrestricted and extensive data for training. Our Ref-L4 benchmark is designed to assess the capabilities of these advanced models without imposing any limitations on the training data sources. The benchmark is divided into two

subsets: a validation set, comprising 30% of the data with 7,231 images, 10,311 instances, and 13,420 referring expressions; and a test set, comprising 70% of the data with 9,467 images, 17,242 instances, and 31,921 referring expressions. Given that our benchmark includes instances from 365 categories, we ensure that each category has at least one sample in both the validation and test sets. While we provide these two splits, we encourage the combined use of both sets for model evaluation, especially in the current LMM era, where the use of unrestricted training data is prevalent.

4. Experiments

Main Result. We evaluate a total of 24 LMMs that can perform the REC task, dividing them into two categories based on their output type: those that produce bounding boxes and those that produce segmentation masks. For models that output segmentation masks, we convert these masks into tight bounding boxes to enable evaluation on our Ref-L4 benchmark. Tab. 5 presents the performance of these models on the validation set, test set, and the combined set, using the metrics defined in Sec. 3.3. The evaluation prompt of GPT-4V is available in Sec. 8.4. Among the models that output bounding boxes, CogVLM-Grounding [63] shows the best performance, while GlaMM [48] leads in performance among the models that output masks.

Scale-Aware Evaluation. In Sec. 3.3, we present a scale-aware evaluation to assess the model’s ability to handle different instance scales. Specifically, we categorize all samples in our benchmark into three sets based on instance size: small, medium, and large. The performance of 24 models is detailed in Tab. 5. Among the bounding-box-output models, SPHINX-v2-1k [29] achieves the best performance with large instances, while CogVLM-Grounding [63] excels with small and medium instances, probably due to its improved visual encoder and supporting higher image resolution than many other works. For mask-output models, GlaMM [48] outperforms all other models across all three sets.

Category-Wise Performance. Each instance in our benchmark is assigned a category label from one of 365 classes. Fig. 5 illustrates the performance of the top four models across these categories, sorted in descending order based on their average per-category performance. The results indicate a training bias issue, as all four models exhibit poor performance on some common categories.

Evaluation on Diverse Data Sources. Our benchmark is derived from COCO and Objects365 datasets. We assess the performance of the top four models with bounding box outputs and the top two models with mask outputs across various subsets originating from either COCO or Objects365. These subsets are: 1) the COCO-derived set (referred to as “COCO”); 2) a subset from Objects365, where the instances have categories that also exist in COCO (referred to as “O365-P1”); 3) another subset from Objects365, where the instances have categories not found in COCO (referred to as “O365-P2”). Fig. 6 presents the performance of these models across the three subsets. The “COCO” set shows higher accuracy compared to the other two sets, partially because most models are trained on the RefCOCO series and have limited exposure to Objects365 images. “O365-P1” exhibits higher accuracy than “O365-P2”, as the latter includes more rare categories.

Extending to Referring Expression Segmentation. The task of Referring Expression Comprehension (REC) can be extended to Referring Expression Segmentation (RES) by predicting a pixel-level mask instead of a bounding box. To extend our Ref-L4 for RES, we use a semi-automated process to transform the bounding boxes into mask annotations. Specifically, for each target instance and its corresponding image, we: 1) input the image and the target instance’s bounding box into the SAM-2 [49] model to generate an initial mask; and 2) manually review and correct the predicted mask if any inaccuracies are identified. We find that SAM-2’s predictions are generally accurate, with only a small proportion of challenging cases (3.5%) requiring manual correction. Tab. 6 presents the evaluation of four models capable of predicting masks. The evaluation protocols remain consistent as above, except that the IoU is

Table 6. Evaluation of four models on the RES benchmark, extended from our Ref-L4 REC benchmark. We merge the validation and test set for evaluation.

Model	mAcc	Acc _{0.5}	Acc _{0.75}	Acc _{0.9}	mAcc-S	mAcc-M	mAcc-L
PixelLM 13B [51]	44.3	67.1	48.2	16.7	20.4	47.8	55.6
LISA [24]	48.6	59.5	50.5	38.4	19.1	53.4	61.3
PSALM [81]	57.4	68.0	60.6	46.3	30.0	64.4	64.8
GlaMM [48]	55.2	66.1	57.9	44.9	20.2	63.6	65.6

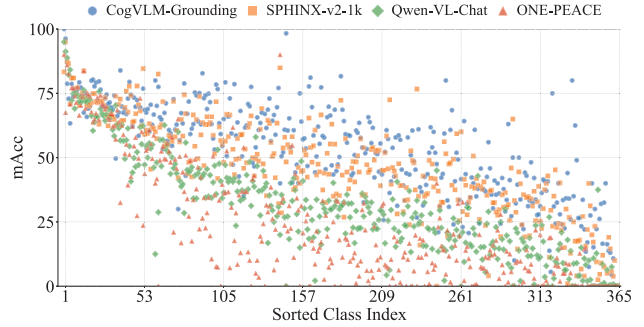


Figure 5. Category-wise performance of the four top-performing models on the val+test set, sorted in descending order based on their average per-category performance. The performance of all models can be found in Sec. 9.1.

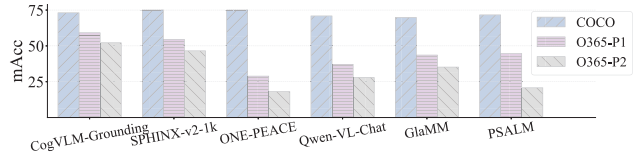


Figure 6. Evaluation of six models on various data sources, with mAcc acting as the metric. The results of all models can be found in Sec. 9.2.

calculated between the predicted mask and the ground-truth mask. In each table, “S”, “M” and “L” represent small, medium and large instances, respectively. In Fig. 15, we provide visualizations of nine randomly selected segmentation annotations from our benchmark.

5. Conclusion

In this work, we first point out several limitations of the current REC benchmarks, such as substantial labeling inaccuracies and very brief referring expressions. To better assess the capabilities of models, particularly those LMMs that can perform the REC task, we present Ref-L4, which features four key characteristics: 1) a large-scale dataset with 45,341 annotations; 2) a wide range of object categories and varying instance scales; 3) detailed referring expressions; and 4) an extensive vocabulary comprising 22,813 unique words. We evaluate a total of 24 models using various evaluation protocols. We wish that Ref-L4 could serve as a valuable resource for researchers and developers, fostering the development of more robust and versatile REC models in the LMM era.

Acknowledgement The work was supported, in part, by Research Grants Council Collaborative Research Fund (under grant number C1045-23G).

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 3, 7, 6
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1
- [3] Yuqi Bu, Liuwu Li, Jiayuan Xie, Qiong Liu, Yi Cai, Qingbao Huang, and Qing Li. Scene-text oriented referring expression comprehension. *IEEE Transactions on Multimedia*, 2022. 3
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3
- [5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 3, 7, 6
- [6] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multi-modal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 7, 6
- [7] 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*, pages 10086–10095, 2020. 3
- [8] Zhihong Chen, Ruifei Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Advancing visual grounding with scene knowledge: Benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15049, 2023. 3
- [9] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1
- [10] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Refer360 degree: A referring expression recognition dataset in 360 degree images. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7189–7202, 2020. 3
- [11] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guess-what?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017. 3
- [12] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1
- [13] Chenyang Gao, Biao Yang, Hao Wang, Mingkun Yang, Wenwen Yu, Yuliang Liu, and Xiang Bai. Textrec: A dataset for referring expression comprehension with reading comprehension. In *International Conference on Document Analysis and Recognition*, pages 402–420. Springer, 2023. 3
- [14] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024. 3, 7, 6
- [15] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 3
- [16] Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. Multi-modal instruction tuned llms with fine-grained visual perception. *arXiv preprint arXiv:2403.02969*, 2024. 1
- [17] Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. Grec: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023. 3
- [18] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *arXiv preprint arXiv:2401.09340*, 2024. 3
- [19] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetrm: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3
- [20] 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. 3
- [21] MARGARITA KOSAREVA. Pushing the limits of visual grounding: Pre-training on large synthetic datasets. *thesis.unipd.it*, 2024. 3
- [22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 3
- [23] Shuhei Kurita, Naoki Katsura, and Eri Onami. Refego: Referring expression comprehension dataset from first-person perception of ego4d. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15214–15224, 2023. 3
- [24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation

- via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 7, 8
- [25] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 3
- [26] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3
- [27] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. Lego: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024. 7, 6
- [28] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3
- [29] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 3, 7, 8, 6
- [30] 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*, pages 23592–23601, 2023. 3
- [31] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017. 3
- [32] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194, 2019. 3
- [33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3
- [34] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. *arXiv preprint arXiv:2404.13013*, 2024. 3
- [35] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1, 3, 4
- [36] AI Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI.*, 2024. 1
- [37] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 1, 3
- [38] OpenAI. Gpt-4 technical report, 2023. 7
- [39] OpenAI. Gpt-4v(ision) system card, 2023.
- [40] OpenAI. Gpt-4v(ision) technical work and authors, 2023. 7
- [41] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 7
- [42] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. *arXiv preprint arXiv:2311.06612*, 2023. 3
- [43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 3
- [44] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks, master of many: Designing general-purpose coarse-to-fine vision-language model. *arXiv preprint arXiv:2312.12423*, 2023. 3
- [45] Lu Qi, Yi-Wen Chen, Lehan Yang, Tiancheng Shen, Xiangtai Li, Weidong Guo, Yu Xu, and Ming-Hsuan Yang. Generalizable entity grounding via assistance of large language model. *arXiv preprint arXiv:2402.02555*, 2024. 3
- [46] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020. 3
- [47] 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. 3
- [48] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 7, 8
- [49] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 8
- [50] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3
- [51] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*, 2023. 7, 8

- [52] Samuel Schuster, Yumin Suh, Konstantinos M Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas, et al. Omnilabel: A challenging benchmark for language-based object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11953–11962, 2023. 3
- [53] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 3, 4
- [54] Haozhan Shen, Tiancheng Zhao, Mingwei Zhu, and Jianwei Yin. Groundvlp: Harnessing zero-shot visual grounding from vision-language pre-training and open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4766–4775, 2024. 3
- [55] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3
- [56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1
- [58] Haiyang Wang, Hao Tang, Li Jiang, Shaoshuai Shi, Muhammad Ferjad Naeem, Hongsheng Li, Bernt Schiele, and Liwei Wang. Git: Towards generalist vision transformer through universal language interface. *arXiv preprint arXiv:2403.09394*, 2024. 1
- [59] Peng Wang, Dongyang Liu, Hui Li, and Qi Wu. Give me something to eat: referring expression comprehension with commonsense knowledge. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 28–36, 2020. 3
- [60] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 2, 7
- [61] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023. 2, 3, 7, 6
- [62] Qinxin Wang, Hao Tan, Sheng Shen, Michael W Mahoney, and Zhewei Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. *arXiv preprint arXiv:2010.05379*, 2020. 3
- [63] Wei Han Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1, 2, 3, 7, 8, 6
- [64] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [65] Wenxuan Wang, Yisi Zhang, Xingjian He, Yichen Yan, Zijia Zhao, Xinlong Wang, and Jing Liu. Beyond literal descriptions: Understanding and locating open-world objects aligned with human intentions. *arXiv preprint arXiv:2402.11265*, 2024. 3
- [66] Yaodong Wang, Zhong Ji, Di Wang, Yanwei Pang, and Xuelong Li. Towards unsupervised referring expression comprehension with visual semantic parsing. *Knowledge-Based Systems*, 285:111318, 2024. 1
- [67] Fei Wei, Xinyu Zhang, Ailing Zhang, Bo Zhang, and Xi-angxiang Chu. Lenna: Language enhanced reasoning detection assistant. *arXiv preprint arXiv:2312.02433*, 2023. 3, 7, 6
- [68] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhansu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 3
- [69] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [70] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023. 3
- [71] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of llms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 4
- [72] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 7
- [73] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 1, 3, 4
- [74] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattrnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 3
- [75] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. *arXiv preprint arXiv:2311.14552*, 2023. 3

- [76] Yufei Zhan, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. *arXiv preprint arXiv:2403.09333*, 2024. 3
- [77] Chao Zhang, Weiming Li, Wanli Ouyang, Qiang Wang, Woo-Shik Kim, and Sunghoon Hong. Referring expression comprehension with semantic visual relationship and word mapping. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1258–1266, 2019. 3
- [78] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 3
- [79] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*, 2023. 3
- [80] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 3
- [81] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. *arXiv preprint arXiv:2403.14598*, 2024. 7, 8
- [82] Haoyu Zhao, Wenhao Ge, and Ying-cong Chen. Llm-optic: Unveiling the capabilities of large language models for universal visual grounding. *arXiv preprint arXiv:2405.17104*, 2024. 1
- [83] Duo Zheng, Tao Kong, Ya Jing, Jiaan Wang, and Xiaojie Wang. Towards unifying reference expression generation and comprehension. *arXiv preprint arXiv:2210.13076*, 2022. 3
- [84] Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6669–6678, 2019. 3
- [85] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 6
- [86] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 3