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

Jierun Chen[*]
HKUST
jcheneh@cse.ust.hk

Fangyun Wei[*†]
Microsoft Research Asia
fawe@microsoft.com

Jinjing Zhao
The University of Sydney
jzha0100@uni.sydney.edu.au

Sizhe Song
HKUST
ssongad@cse.ust.hk

Bohuai Wu
HKUST
bwual@cse.ust.hk

Zhuoxuan Peng
HKUST
zpengac@cse.ust.hk

S.-H. Gary Chan
HKUST
gchan@cse.ust.hk

Hongyang Zhang
University of Waterloo
hongyang.zhang@uwaterloo.ca

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 scales 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) [47, 38, 81, 21, 43, 75, 65] 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 (LLMs) [55, 56, 37, 9, 15, 1, 28, 19, 26]. This progress is particularly evident in the exceptional performance of large multimodal models (LMMs) [62, 31, 13, 60, 2, 5, 66, 16, 57, 17, 80] on well-known benchmarks such as RefCOCO [71], RefCOCO+ [71], and RefCOCOg [36]. These models

[*]Equal contribution.
[†]Corresponding author.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

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 [71]	21,586	3,054	14%
RefCOCO+ [71]	21,373	5,201	24%
RefCOCOg [36]	14,498	675	5%

Table 2: The performance of four LMMs capable of handling the REC task on both the cleaned and original versions of the 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† [60]	OFA-L† [59]	OFA-L [59]	Qwen-VL [2]	CogVLM-Grounding [62]
RefCOCO [71]	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+ [71]	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 [36]	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)**

have demonstrated remarkable accuracy, with CogVLM [62], for instance, achieving an impressive accuracy rate of 92.44% on the RefCOCO benchmark.

This paper begins with a critical question: do existing REC benchmarks truly capture the comprehensive capabilities of LMMs? The foundational benchmarks, RefCOCO [71], RefCOCO+ [71], and RefCOCOg [36], were introduced sequentially in 2015, 2016, and 2016, respectively. In RefCOCO, the referring expressions are notably succinct, ranging from single words like "*lady*" and "*yellow*" 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 "*plastic cup with just ice*" and "*man on screen*". Conversely, RefCOCOg provides more elaborate annotations, including examples such as "*a table of food, with plates, a pizza, pitchers, and glasses*" and "*a red and white checkered table with two wooden chairs*". These variations highlight the evolution and complexity of referring expressions across different benchmarks, raising the question of whether they can effectively assess the nuanced capabilities of modern LMMs in understanding diverse linguistic inputs and associating languages with visual elements.

Labeling Error Rates of Existing Benchmarks. To begin, 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 Section A. As illustrated in Table 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 [60], OFA-L [59], Qwen-VL [2], and CogVLM-Grounding [62]—on both the cleaned and original versions of these datasets, as shown in Table 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 impacted the models' true capabilities. *To support further research in the REC field, we release the cleaned versions of RefCOCO, RefCOCO+, and RefCOCOg.*

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

Table 3: Comparison between our Ref-L4 benchmark and other REC benchmarks, including RefCOCO [71], RefCOCO+ [71], and RefCOCOg [36]. 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	Image Size	Vocab.
RefCOCO [71]	3,000	7,596	21,586	71	3.6	105 - 607	230 - 640	3,525
RefCOCO+ [71]	3,000	7,578	21,373	71	3.6	105 - 607	230 - 640	4,387
RefCOCOg [36]	3,900	7,596	14,498	78	8.4	83 - 610	277 - 640	5,050
Ref-L4 (Ours)	9,735	18,653	45,341	365	24.2	30 - 3,767	230 - 6,606	22,813

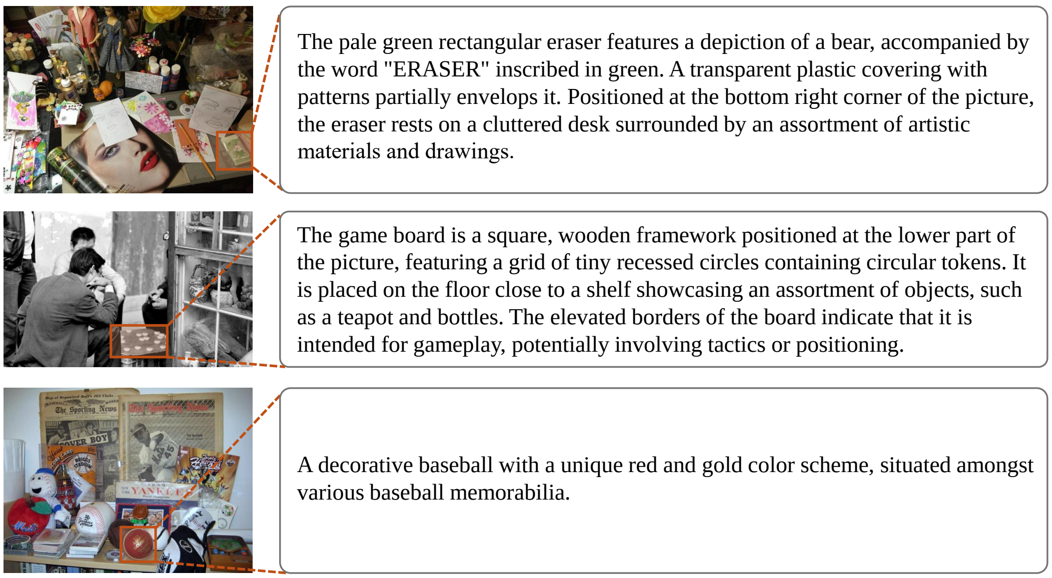


Figure 1: Examples from our Ref-L4 benchmark. We offer a detailed referring expression for each target instance represented by a bounding box. Zoom in for better visualization.

make Ref-L4 a comprehensive benchmark for assessing the REC capabilities of contemporary LMMs. Table 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 scales, from 30 to $3,767$, measured by the square root of the instance 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 Figure 1.
- *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

$Acc_{0.5}$ to $Acc_{0.9}$ 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 are available at* `https://github.com/JierunChen/Ref-L4`.

2 Related Work

REC and Its Benchmarks. Referring Expression Comprehension (REC) [47, 38, 81, 21, 43, 75] is a task that involves identifying a specific object within an image based on a given referring expression. Unlike object detection [30, 23, 52, 50, 4], which operates within fixed categories and a single visual modality, REC necessitates understanding free-form text to locate objects of any category. Phrase Grounding [44, 67, 14, 34, 27, 76, 61] 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. Current REC benchmarks such as RefCOCO [71], RefCOCO+[71], and RefCOCOg[36] 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 [33, 8, 48, 7, 64, 24, 58, 10, 3, 12, 11, 18] are often designed for specific scenarios. For example, CLEVR-Ref+[33] focuses on simple objects like boxes, spheres, and cylinders. SK-VG[8] integrates prior scene knowledge as additional input, while RefCrowd [48] 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 [20, 72, 32, 54, 82, 68, 83] to generalist models or large multimodal models (LMMs)[62, 31, 13, 60, 2, 5, 66, 78, 73, 74, 45, 77, 63, 53, 35, 46, 22]. Notable examples of these LMMs include CogVLM-Grounding[62], SPHINX [31, 13], ONE-PEACE [60], Qwen-VL-Chat [2], MiniGPTv2 [5], and Lenna [66]. 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 [71], RefCOCO+ [71], and RefCOCOg [36] datasets; and 2) the test set from the large-scale object detection dataset Objects365 [52]. 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 bea". 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

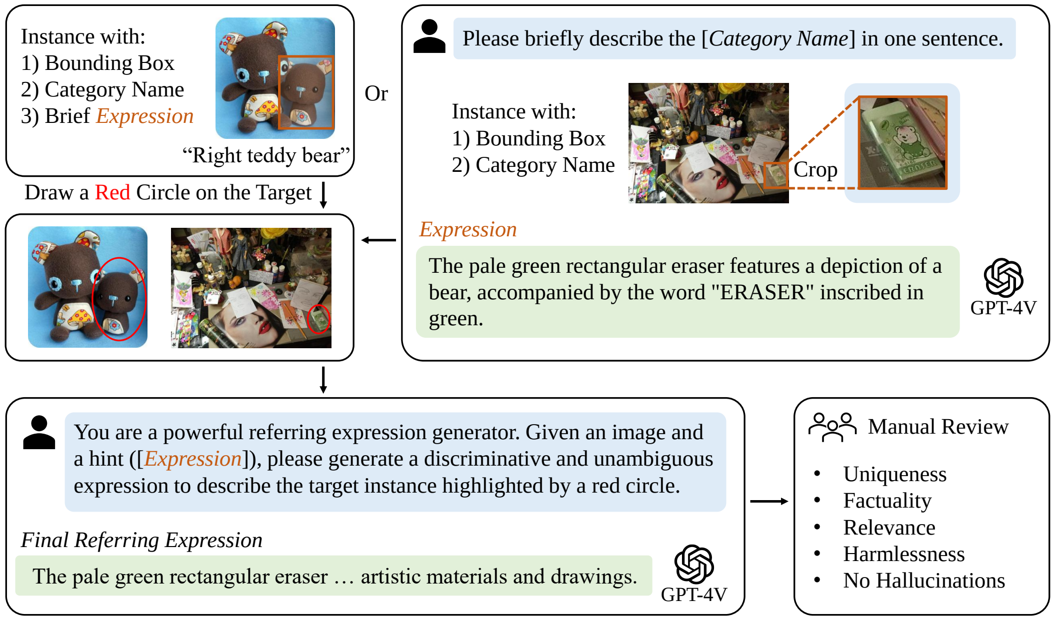


Figure 2: Pipeline of generating a referring expression for a target instance.

$N = \min(35, \text{the number of instances for the specific class})$; 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.

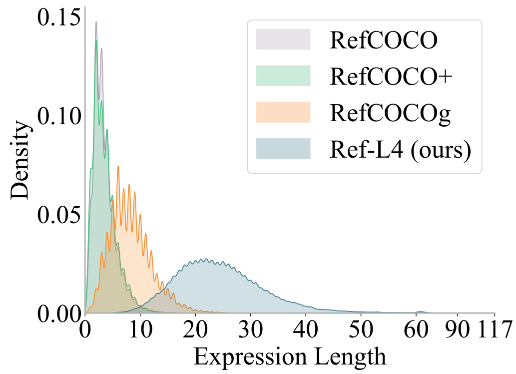
Referring Expression Generation. Given a target instance and its corresponding image, we leverage GPT-4V with human reviewers in the loop to generate its precise and detailed referring expressions. Figure 2 illustrates the three-step generation process:

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 Section B.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.

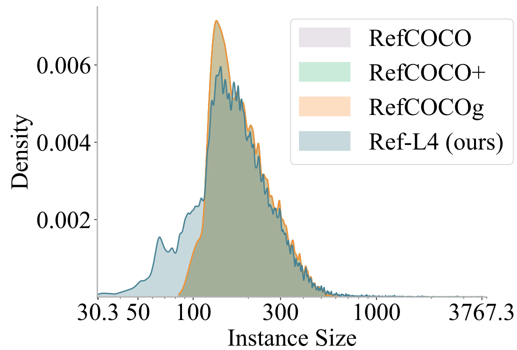
Step-2: Drawing inspiration from recent studies on GPT-4V [69], where GPT-4V is able to pay more attention to instances highlighted by a red circle within an image, we similarly encircle the target instance in red to facilitate GPT-4V in generating a context-aware referring expression. Following this, as depicted in Figure 2, we process the image and use the prompt outlined in Section B.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.

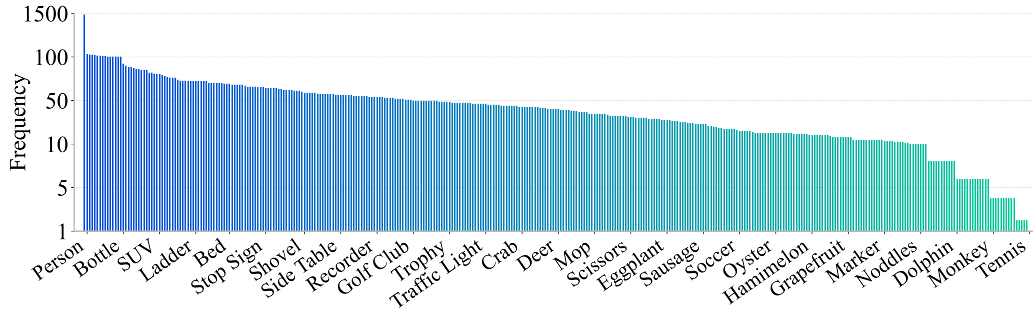
Annotation Expansion. To date, we have compiled 18,653 unique referring expressions, each describing a distinct instance. To assess the robustness of REC models to diverse language inputs, we employ a two-stage rephrasing process to expand our benchmark: 1) utilizing GPT-4 with the prompt detailed in Section B.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. Consequently, our final Ref-L4 benchmark encompasses 9,735 images with 45,341 referring expressions, each accurately describing one of the 18,653 unique instances.



(a) The distribution of expression length.

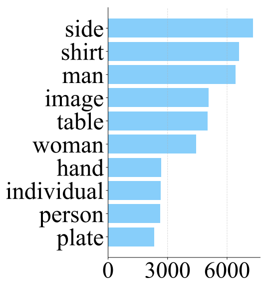


(b) The distribution of instance size.

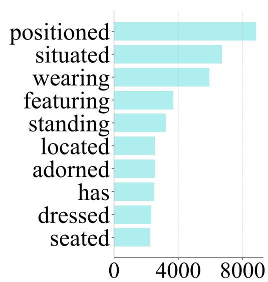


(c) The distribution of instance numbers over 365 categories.

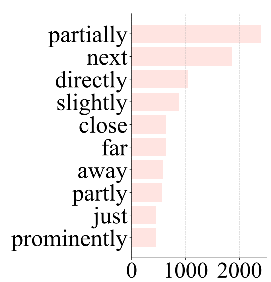
Figure 3: Analysis of referring expression length, instance size, and category distribution.



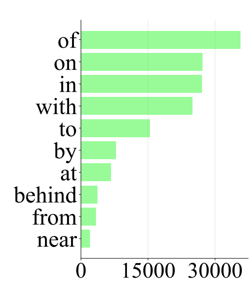
(a) Nouns.



(b) Verbs.



(c) Adverbs.



(d) Prepositions.

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

3.2 Analysis

Expression Length. Figure 3a illustrates the distribution of expression lengths across four different datasets: RefCOCO, RefCOCO+, RefCOCOg, and our Ref-L4. Due the 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 designed to push the boundaries of current REC models, requiring them to process and comprehend more intricate and detailed descriptions.

Instance Size. In Figure 3b, we present a density plot comparing the instance sizes across four benchmarks. We define the instance size as the square root of the normalized size, $\sqrt{(hw)/(HW)}$, where (h, w) represents the dimensions of the instance and (H, W) represents the dimensions of the image. All benchmarks exhibit a peak density around an instance size of 160. Our Ref-L4 benchmark shows a wider distribution range compared to the other three, indicating that our Ref-L4 captures a broader spectrum of instance sizes.

Categories. Our Ref-4L 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. Figure 3c presents the distribution 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".

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. Figure 4 illustrates the 10 most frequently used nouns, verbs, adverbs, and prepositions.

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 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 $\text{Acc}_{0.9}$ 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. Table 4 presents the performance of these models on the validation set, test set, and the combined set, using the metrics defined in Section 3.3. The evaluation prompt of GPT-4V is available in Section B.4. Among the models that output bounding boxes, CogVLM-Grounding [62] shows the best performance, while GlaMM [49] leads in performance among the models that output masks.

Category-Wise Performance. Each instance in our benchmark is assigned a category label from one of 365 classes. Figure 5 illustrates the performance of the top four models across these categories,

Table 4: 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
	$Acc_{0.5}$	$Acc_{0.75}$	$Acc_{0.9}$	mAcc	mAcc	mAcc
GPT-4V [39–41]	9.91	1.19	0.12	2.88	2.96	2.85
KOSMOS-2 [42]	48.53	38.34	17.54	34.72	34.89	34.64
OFA-Tiny [59]	55.21	43.22	27.70	41.44	41.53	41.40
OFA-Large [59]	72.53	62.31	45.02	59.17	59.42	59.07
Ferret-7b [70]	57.54	42.44	21.01	40.29	40.31	40.28
Ferret-13b [70]	64.44	49.04	27.46	46.88	47.31	46.71
GroundingGPT [29]	60.84	40.48	12.00	38.19	38.42	38.09
Shikra-7b [6]	65.06	39.62	10.45	38.60	38.91	38.47
Lenna [66]	65.90	58.55	45.58	55.69	55.88	55.60
MiniGPTv2 [5]	66.93	50.50	25.30	47.15	47.43	47.03
Qwen-VL-Chat [2]	73.80	58.05	37.16	55.94	56.18	55.83
ONE-PEACE [60]	70.82	60.09	36.12	55.07	55.49	54.89
SPHINX-MoE [13]	66.23	44.90	15.32	42.38	42.80	42.21
SPHINX-MoE-1k [13]	74.45	62.70	38.85	58.07	58.35	57.95
SPHINX [31]	74.78	53.65	21.15	50.09	50.33	49.99
SPHINX-1k [31]	78.52	62.17	32.95	57.57	57.91	57.42
SPHINX-v2-1k [31]	81.31	70.49	46.59	65.39	65.67	65.27
CogVLM-Grounding [62]	**81.70**	**70.77**	**48.35**	**66.09**	**66.25**	**66.02**
PixelLM-7B† [51]	41.83	27.57	13.32	27.10	27.09	27.11
PixelLM-13B† [51]	49.89	35.37	18.42	34.10	34.52	33.92
LISA-Explanatory† [25]	65.12	52.35	38.26	50.77	50.89	50.72
LISA† [25]	66.23	54.02	39.73	52.18	52.44	52.07
PSALM† [79]	67.26	58.22	44.11	55.46	55.68	55.37
GlaMM† [49]	**71.90**	**60.27**	**45.15**	**57.89**	**58.16**	**57.78**

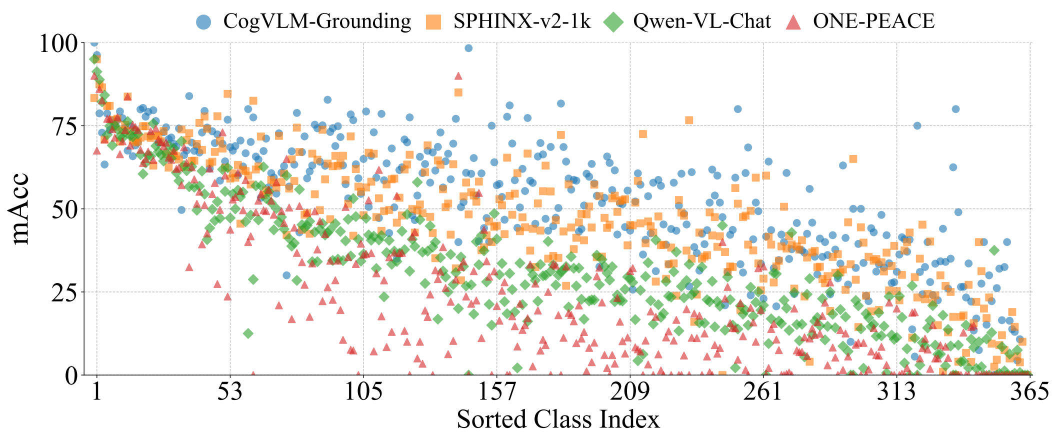


Figure 5: Category-wise performance of the four top-performing models on the val+test set, sorted in descending order based on their average per-category performance. The performance of all models can be found in Section C.1.

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.

Scale-Aware Evaluation. In Section 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 Table 5. Among the bounding-box-output models, CogVLM-Grounding [62] excels with small and medium instances, while SPHINX-v2-1k [31] achieves the best performance with large instances. For mask-output models, GlaMM [49] outperforms all other models across all three sets.

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

Table 5: Scale-aware evaluation across 24 models on our Ref-L4 benchmark.

Model	Small Size		Medium Size		Large Size	
	$Acc_{0.5}$	mAcc	$Acc_{0.5}$	mAcc	$Acc_{0.5}$	mAcc
GPT-4V [39–41]	2.13	0.49	10.29	2.78	14.93	4.83
KOSMOS-2 [42]	24.19	11.63	46.95	32.91	69.32	54.98
OFA-Tiny [59]	17.91	11.49	65.13	49.00	64.46	49.61
OFA-Large [59]	40.13	27.07	81.03	66.49	80.78	69.36
Ferret-7b [70]	30.93	14.57	62.40	43.72	68.18	52.92
Ferret-13b [70]	36.46	17.88	70.50	51.86	73.92	59.09
GroundingGPT [29]	24.43	10.28	67.67	41.04	75.09	53.47
Shikra-7b [6]	43.91	18.50	75.98	46.27	60.60	39.34
Lenna [66]	31.02	23.48	72.90	61.53	78.72	68.66
MiniGPTv2 [5]	32.99	14.85	73.67	51.16	79.52	63.53
Qwen-VL-Chat [2]	47.66	26.26	79.80	61.06	82.01	68.37
ONE-PEACE [60]	22.18	13.98	83.26	63.39	83.81	70.04
SPHINX-MoE [13]	39.48	16.39	72.97	46.38	73.55	54.17
SPHINX-MoE-1k [13]	58.96	37.61	77.80	61.53	79.70	66.77
SPHINX [31]	48.82	22.08	80.56	54.10	83.27	63.34
SPHINX-1k [31]	59.48	33.21	82.95	61.82	84.40	67.68
SPHINX-v2-1k [31]	65.23	43.43	84.00	68.45	**88.21**	**75.91**
CogVLM-Grounding [62]	**75.06**	**52.85**	**86.43**	**71.31**	77.91	66.25
PixelLM-7B† [51]	8.25	4.05	43.90	27.33	62.72	43.64
PixelLM-13B† [51]	17.05	8.54	53.40	35.48	67.59	50.34
LISA-Explanatory† [25]	39.11	27.16	70.03	54.61	75.25	61.09
LISA† [25]	39.24	27.49	71.17	56.05	77.01	63.22
PSALM† [79]	37.35	28.43	75.06	61.79	74.97	63.74
GlaMM† [49]	**47.07**	**34.36**	**77.17**	**62.28**	**80.50**	**67.14**

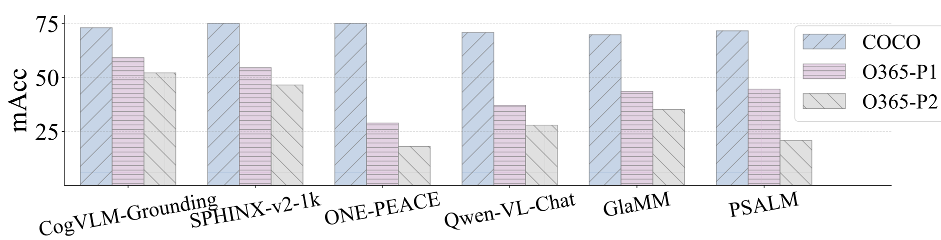


Figure 6: Evaluation of six models on various data sources, with mAcc acting as the metric. The results of all models can be found in Section C.2.

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"). Figure 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.

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.

References

[1] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[3] Y. Bu, L. Li, J. Xie, Q. Liu, Y. Cai, Q. Huang, and Q. Li. Scene-text oriented referring expression comprehension. *IEEE Transactions on Multimedia*, 2022.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[5] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[6] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[7] Z. Chen, P. Wang, L. Ma, K.-Y. K. Wong, and Q. Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10086–10095, 2020.

[8] Z. Chen, R. Zhang, Y. Song, X. Wan, and G. 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.

[9] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[10] V. Cirik, T. Berg-Kirkpatrick, and L.-P. 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.

[11] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

[12] C. Gao, B. Yang, H. Wang, M. Yang, W. Yu, Y. Liu, and X. Bai. Textrec: A dataset for referring expression comprehension with reading comprehension. In *International Conference on Document Analysis and Recognition*, pages 402–420. Springer, 2023.

[13] P. Gao, R. Zhang, C. Liu, L. Qiu, S. Huang, W. Lin, S. Zhao, S. Geng, Z. Lin, P. Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.

[14] A. Gupta, P. Dollar, and R. 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.

[15] Y. Hao, H. Song, L. Dong, S. Huang, Z. Chi, W. Wang, S. Ma, and F. Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022.

[16] J. He, Y. Wang, L. Wang, H. Lu, J.-Y. He, J.-P. Lan, B. Luo, and X. Xie. Multi-modal instruction tuned llms with fine-grained visual perception. *arXiv preprint arXiv:2403.02969*, 2024.

[17] Z. Huang, Z. Zhang, Z.-J. Zha, Y. Lu, and B. Guo. Relationvlm: Making large vision-language models understand visual relations. *arXiv preprint arXiv:2403.12801*, 2024.

[18] B. Jia, Y. Chen, H. Yu, Y. Wang, X. Niu, T. Liu, Q. Li, and S. Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *arXiv preprint arXiv:2401.09340*, 2024.

[19] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[20] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[21] S. Kazemzadeh, V. Ordonez, M. Matten, and T. 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.

[22] M. KOSAREVA. Pushing the limits of visual grounding: Pre-training on large synthetic datasets.

[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[24] S. Kurita, N. Katsura, and E. 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.

[25] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[26] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[27] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. 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.

[28] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

[29] Z. Li, Q. Xu, D. Zhang, H. Song, Y. Cai, Q. Qi, R. Zhou, J. Pan, Z. Li, V. T. Vu, et al. Lego: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. 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.

[31] Z. Lin, C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. 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.

[32] J. Liu, L. Wang, and M.-H. Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017.

[33] R. Liu, C. Liu, Y. Bai, and A. 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.

[34] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[35] C. Ma, Y. Jiang, J. Wu, Z. Yuan, and X. Qi. Groma: Localized visual tokenization for grounding multimodal large language models. *arXiv preprint arXiv:2404.13013*, 2024.

[36] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[37] A. Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI.*, 2024.

[38] V. K. Nagaraja, V. I. Morariu, and L. 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.

[39] OpenAI. Gpt-4 technical report, 2023.

[40] OpenAI. Gpt-4v(ision) system card. 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

[41] OpenAI. Gpt-4v(ision) technical work and authors. 2023. URL https://cdn.openai.com/contributions/gpt-4v.pdf.

[42] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[43] R. Pi, L. Yao, J. Gao, J. Zhang, and T. Zhang. Perceptiongpt: Effectively fusing visual perception into llm. *arXiv preprint arXiv:2311.06612*, 2023.

[44] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. 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.

[45] S. Pramanick, G. Han, R. Hou, S. Nag, S.-N. Lim, N. Ballas, Q. Wang, R. Chellappa, and A. Almahairi. Jack of all tasks, master of many: Designing general-purpose coarse-to-fine vision-language model. *arXiv preprint arXiv:2312.12423*, 2023.

[46] L. Qi, Y.-W. Chen, L. Yang, T. Shen, X. Li, W. Guo, Y. Xu, and M.-H. Yang. Generalizable entity grounding via assistance of large language model. *arXiv preprint arXiv:2402.02555*, 2024.

[47] Y. Qiao, C. Deng, and Q. Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020.

[48] H. Qiu, H. Li, T. Zhao, L. Wang, Q. Wu, and F. Meng. Refcrowd: Grounding the target in crowd with referring expressions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4435–4444, 2022.

[49] H. Rasheed, M. Maaz, S. Shaji, A. Shaker, S. Khan, H. Cholakkal, R. M. Anwer, E. Xing, M.-H. Yang, and F. S. Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023.

[50] J. Redmon, S. Divvala, R. Girshick, and A. 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.

[51] Z. Ren, Z. Huang, Y. Wei, Y. Zhao, D. Fu, J. Feng, and X. Jin. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*, 2023.

[52] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. 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.

[53] H. Shen, T. Zhao, M. Zhu, and J. 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*, volume 38, pages 4766–4775, 2024.

[54] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[55] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[56] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[57] H. Wang, H. Tang, L. Jiang, S. Shi, M. F. Naeem, H. Li, B. Schiele, and L. Wang. Git: Towards generalist vision transformer through universal language interface. *arXiv preprint arXiv:2403.09394*, 2024.

[58] P. Wang, D. Liu, H. Li, and Q. 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.

[59] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. 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.

[60] P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, and C. Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023.

[61] Q. Wang, H. Tan, S. Shen, M. W. Mahoney, and Z. Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. *arXiv preprint arXiv:2010.05379*, 2020.

[62] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[63] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. 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.

[64] W. Wang, Y. Zhang, X. He, Y. Yan, Z. Zhao, X. Wang, and J. Liu. Beyond literal descriptions: Understanding and locating open-world objects aligned with human intentions. *arXiv preprint arXiv:2402.11265*, 2024.

[65] Y. Wang, Z. Ji, D. Wang, Y. Pang, and X. Li. Towards unsupervised referring expression comprehension with visual semantic parsing. *Knowledge-Based Systems*, 285:111318, 2024.

[66] F. Wei, X. Zhang, A. Zhang, B. Zhang, and X. Chu. Lenna: Language enhanced reasoning detection assistant. *arXiv preprint arXiv:2312.02433*, 2023.

[67] C. Wu, Z. Lin, S. Cohen, T. Bui, and S. 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.

[68] B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan, and H. 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.

[69] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

[70] H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, Z. Wang, L. Cao, S.-F. Chang, and Y. Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

[71] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. 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.

[72] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

[73] Y. Zhan, Y. Zhu, Z. Chen, F. Yang, M. Tang, and J. Wang. Griffon: Spelling out all object locations at any granularity with large language models. *arXiv preprint arXiv:2311.14552*, 2023.

[74] Y. Zhan, Y. Zhu, H. Zhao, F. Yang, M. Tang, and J. Wang. Griffon v2: Advancing multimodal perception with high-resolution scaling and visual-language co-referring. *arXiv preprint arXiv:2403.09333*, 2024.

[75] C. Zhang, W. Li, W. Ouyang, Q. Wang, W.-S. Kim, and S. 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.

[76] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022.

[77] H. Zhang, H. Li, F. Li, T. Ren, X. Zou, S. Liu, S. Huang, J. Gao, L. Zhang, C. Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*, 2023.

[78] H. Zhang, H. You, P. Dufter, B. Zhang, C. Chen, H.-Y. Chen, T.-J. Fu, W. Y. Wang, S.-F. Chang, Z. Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024.

[79] Z. Zhang, Y. Ma, E. Zhang, and X. Bai. Psalm: Pixelwise segmentation with large multi-modal model. *arXiv preprint arXiv:2403.14598*, 2024.

[80] H. Zhao, W. Ge, and Y.-c. Chen. Llm-optic: Unveiling the capabilities of large language models for universal visual grounding. *arXiv preprint arXiv:2405.17104*, 2024.

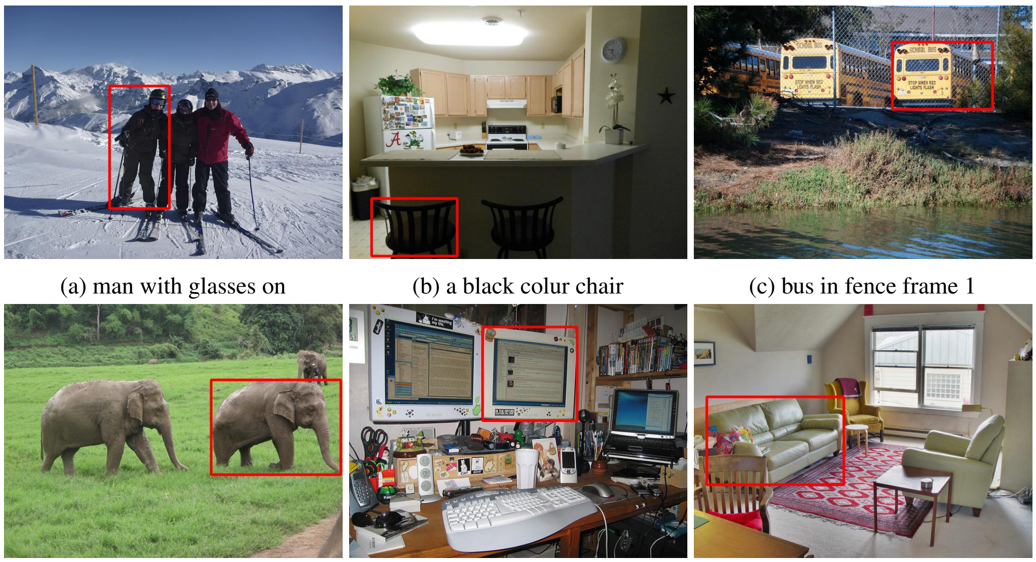
[81] D. Zheng, T. Kong, Y. Jing, J. Wang, and X. Wang. Towards unifying reference expression generation and comprehension. *arXiv preprint arXiv:2210.13076*, 2022.

[82] Z. Zheng, W. Wang, S. Qi, and S.-C. 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.

[83] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.

A Labeling Errors in Existing Benchmarks

In the REC task, a referring expression should uniquely describe an instance, which is represented by an accurate bounding box. We have identified and visualized three common types of labeling errors in the RefCOCO, RefCOCO+, and RefCOCOg benchmarks: 1) non-unique referring expressions (Figure 7), which refer to multiple instances within the same image; 2) inaccurate bounding boxes (Figure 8); and 3) misalignment between target instances and their referring expressions (Figure 9), where the referring expressions are either ambiguous or do not refer to any instance in the image.



(a) man with glasses on (b) a black colur chair (c) bus in fence frame 1

(d) an elephant walking in the grass (e) a white computer screen (f) white couch

Figure 7: Visualization of labeling errors, where a referring expression refers to multiple instances within the same image. For each sub-figure, we display the original bounding box annotation with a red rectangle and include the corresponding referring expression in the caption.

B Prompts

B.1 Prompt for Context-Independent Description Generation

Briefly describe the [*Category Name*] in one sentence. Begin your description with the object name, including adjectives if appropriate to describe its color or shape. Focus only on visible features and avoid mentioning blurriness.

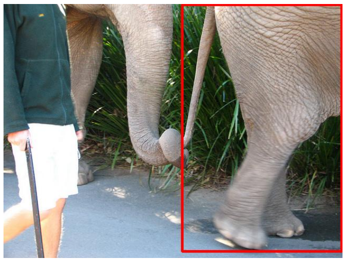
Input image: [*Cropped Image*].

B.2 Prompt for Context-Aware Description Generation

You are a sophisticated referring expression generator. Your task is to generate a clear and specific description for the target instance highlighted by a red circle in the provided image, based on a given hint and the following criteria:

Criteria 1: The description should enable individuals to understand and accurately identify the specified region within the image.

Criteria 2: The description may should various attributes such as category, shape, size, color, visibility, exposure, texture, orientation, absolute position, relative position, facial features, clothing, accessories, gestures, context, semantic attributes, emotions, age, gender, posture, action, and especially interactions with other instances. The selection of features should be relevant to the particular region and the image context.



(a) tail of elephant

(b) red jacket

(c) name on oven

(d) pumpkin

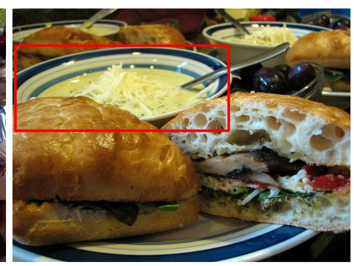
(e) a knife cutting a cake

(f) white hair

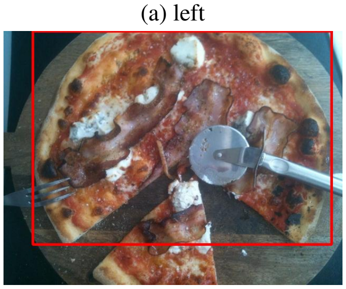
Figure 8: Visualization of labeling errors, where the bounding box annotations are inaccurate. For each sub-figure, we display the original bounding box annotation with a red rectangle and include the corresponding referring expression in the caption.

(a) left

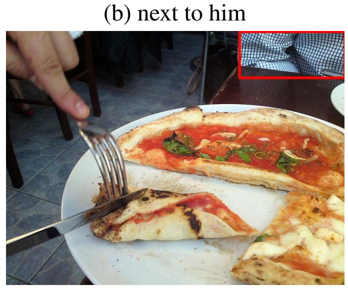
(b) next to him



(c) yep



(d) not the slice



(e) why is the game doing this checker shirt

(f) last hotdog mess thing

Figure 9: Visualization of labeling errors, where the referring expressions are either ambiguous or do not refer to any instance in the image. For each sub-figure, we display the original bounding box annotation with a red rectangle and include the corresponding referring expression in the caption.

Criteria 3: The red circle is solely for highlighting the region of interest. Do not refer to it in your descriptions.

Criteria 4: Avoid using unnecessary words like “look for”, “spot”, “observe”, “find”, “notice”, “identify”, “outline”, “target” and “question”.

Criteria 5: Ensure that the subject of each sentence matches the subject given in the hints. Do not incorrectly use the subject as the object.

Criteria 6: Use the correct singular or plural form when referring to the target, which may be a single object, a pair of objects, or a group of objects.

Criteria 7: Integrate all relevant information from the hints, noting that some hints may be redundant or contain errors.

Input image: [*Raw Image*].

Hint: [*Context-Independent Description*].

B.3 Prompt for Rephrasing Referring Expressions

Rewrite the subsequent description while preserving the main information. Utilize varied expressions and reorganize the sentences if necessary. Begin each sentence with the same subject being referred to.

Description: [*The Referring Expression to be Rephrased*].

B.4 Prompt for GPT4-V Evaluation

You are an expert in referring expression comprehension and localization. Your task is to locate the object in the image based on the provided expression. The coordinates range from the top left (0, 0) to the bottom right ([*Image Width*], [*Image Height*]). Please provide the bounding box in the format (x_0, y_0, x_1, y_1), where (x_0, y_0) represents the top-left corner and (x_1, y_1) represents the bottom-right corner.

Expression: [*The Referring Expression*].

C More Experiments

C.1 Category-Wise Performance.

Figure 5 presents the per-category performance of the top four models. In Figures 10 and 11, we show the performance for all 24 models on a per-category basis, with mAcc serving as the metric, along with the average performance for each model across all categories.

C.2 Evaluation on Diverse Data Sources.

Figure 6 illustrates the performance of six models across three subsets, namely “COCO”, “O365-P1” and “O365-P2”. In Figure 12, the comprehensive results of 24 models across the same three subsets are displayed.

D Limitations and Broad Impacts

Ref-L4 provides a more comprehensive and detailed evaluation of REC capabilities, helping to better understand and improve the performance of large multimodal models capable of handling the REC task. The public availability of Ref-L4 and its evaluation code encourages further research and collaboration, driving innovation and advancements in the field of REC and beyond. While Ref-L4 aims to cover a wide range of scenarios, it may still miss out on specific edge cases or unique contexts that could be encountered in real-world applications. The detailed and lengthy referring expressions might pose a challenge for current models, requiring significant advancements in natural language processing and comprehension capabilities.

E Author Statement

The authors of the Ref-L4 benchmark accept full accountability for any rights violations, such as copyright infringement or other legal breaches. They emphasize that all data included in the Ref-L4 dataset adheres to the licensing agreements of the original source datasets. The Ref-L4 benchmark is made available under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. Meticulous attention has been paid to ensure that the dataset upholds the highest legal and ethical standards. The authors are committed to addressing any issues arising from the use of this dataset and stand prepared to take necessary actions to resolve them.

F Maintenance and Long Term Preservation

To ensure the benchmark remains relevant and useful for evaluating REC models, we will establish a protocol for regular updates. This includes the addition of new image sets and text annotations that reflect current trends and challenges in the field. A version control system will be implemented to track changes and updates to the benchmark. Each version will be documented with detailed notes on the modifications, including the addition of new data, changes to annotation guidelines, and improvements based on user feedback. We will utilize reliable cloud storage solutions with multiple redundancy mechanisms to safeguard against data loss.

G Datasheet

The datasheet of our Ref-L4 benchmark can be found in the supplementary material.

H Links and Licenses

Evaluation Code. The evaluation code is available at `https://github.com/JierunChen/Ref-L4`.

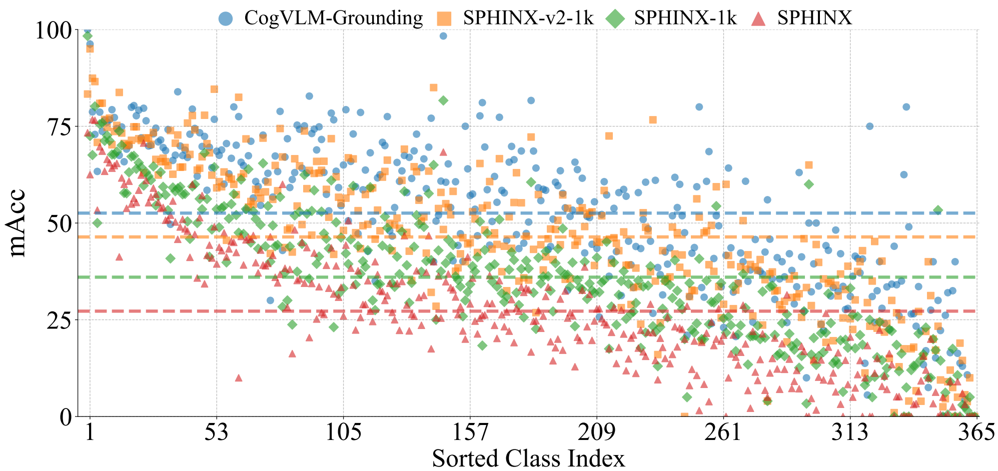
Benchmark Link. Ref-L4 is available for download from the Huggingface platform at `https://huggingface.co/datasets/JierunChen/Ref-L4`.

Croissant Metadata. The croissant format metadata for Ref-L4 can be accessed at `https://huggingface.co/api/datasets/JierunChen/Ref-L4/croissant`.

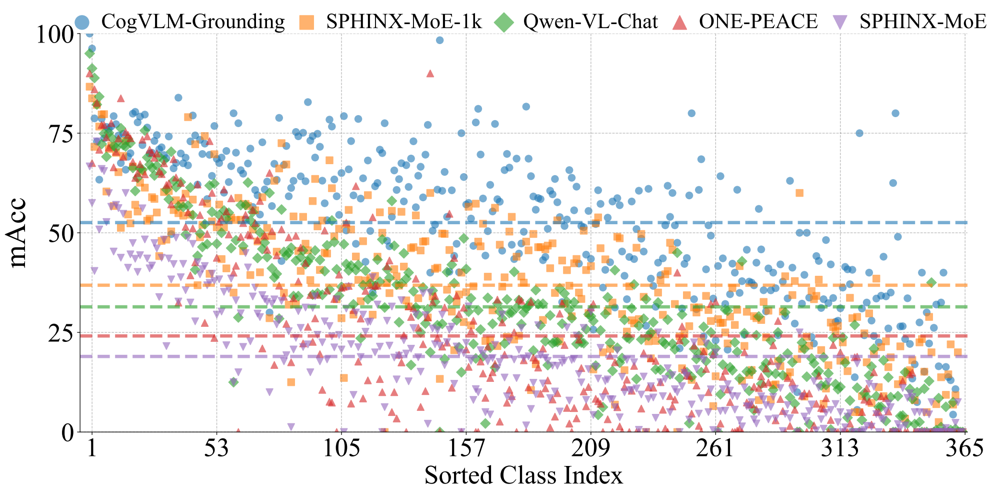
DOI. The DOI of Ref-L4 is 10.57967/hf/2388.

License. Ref-L4 is distributed under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. The images included in Ref-L4 are derived from the following sources, each governed by their respective licenses:

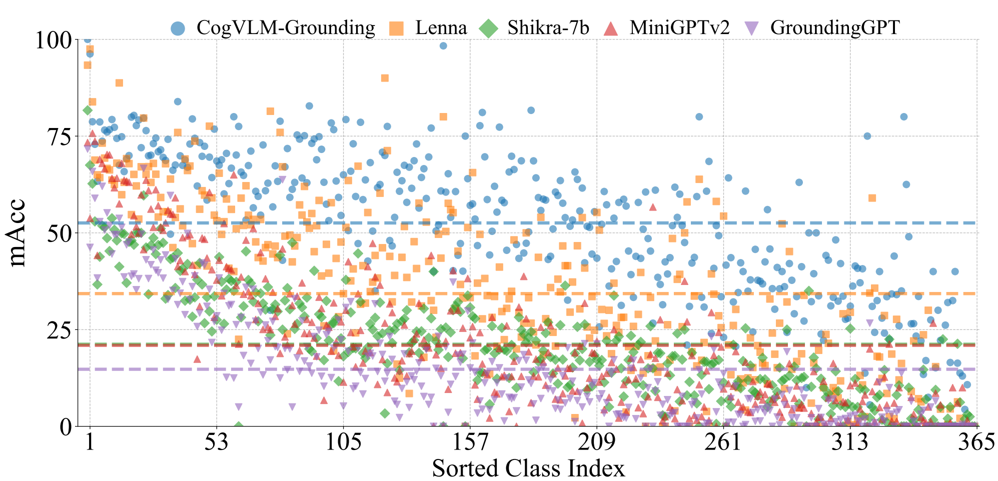
- RefCOCO: Licensed under the Apache-2.0 license.
- RefCOCO+: Licensed under the Apache-2.0 license.
- RefCOCOg: Licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.
- COCO 2014: Licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.
- Objects365: Licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.



(a) The average performance across all categories (dot lines) for CogVLM-Grounding [62], SPHINX-v2-1k [31], SPHINX-1k [31], and SPHINX1 [31] are 52.56, 46.40, 36.01, and 26.95, respectively.

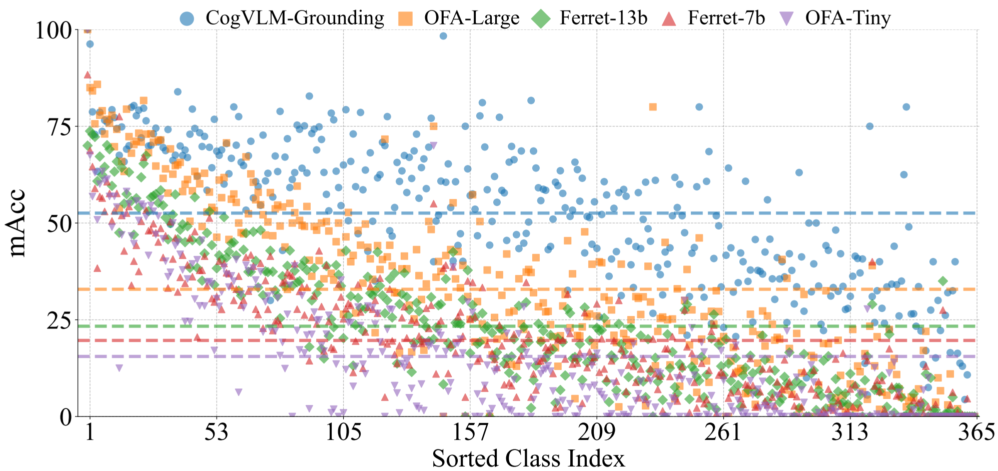


(b) The average performance across all categories (dot lines) for SPHINX-MoE-1k [13], Qwen-VL-Chat [2], ONE-PEACE [60], and SPHINX-MoE [13] are 36.84, 31.41, 24.11, and 18.77, respectively.

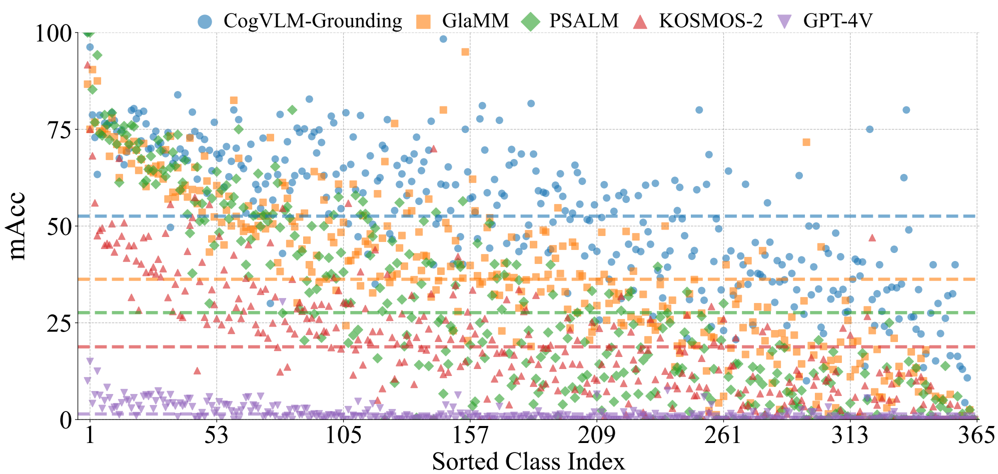


(c) The average performance across all categories (dot lines) for Lenna [66], Shikra-7b [6], MiniGPTv2 [5], and GroundingGPT [29] are 34.30, 21.22, 21.13, and 14.60, respectively.

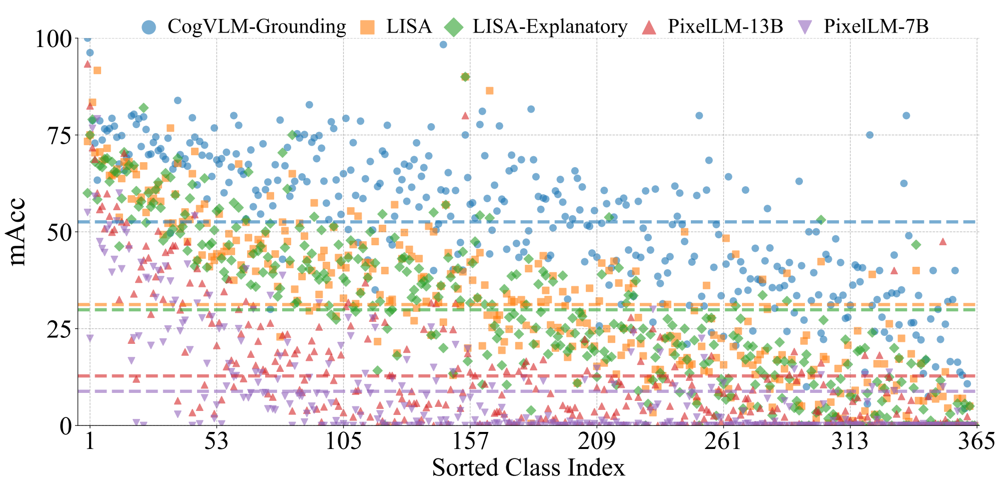
Figure 10: Category-wise performance of 24 models (part-1), sorted in the same order as in Figure 5. We use CogVLM-Grounding as a reference for comparison in each sub-figure.



(a) The average performance across all categories (dot lines) for OFA-Large [59], Ferret-13b [70], Ferret-7b [70] and OFA-Tiny [59] are 32.88, 23.33, 20.27, and 15.37, respectively.



(b) The average performance across all categories (dot lines) for GlaMM [49], PSALM [79], KOSMOS-2 [42] and GPT-4V [39–41] are 36.25, 27.62, 19.37, and 1.42, respectively.



(c) The average performance across all categories (dot lines) for LISA [25], LISA-Explanatory [25], PixelLM-13B [51] and PixelLM-7B [51] are 31.22, 29.87, 13.19, and 8.74, respectively.

Figure 11: Category-wise performance of 24 models (part-2), sorted in the same order as in Figure 5. We use CogVLM-Grounding as a reference for comparison in each sub-figure.

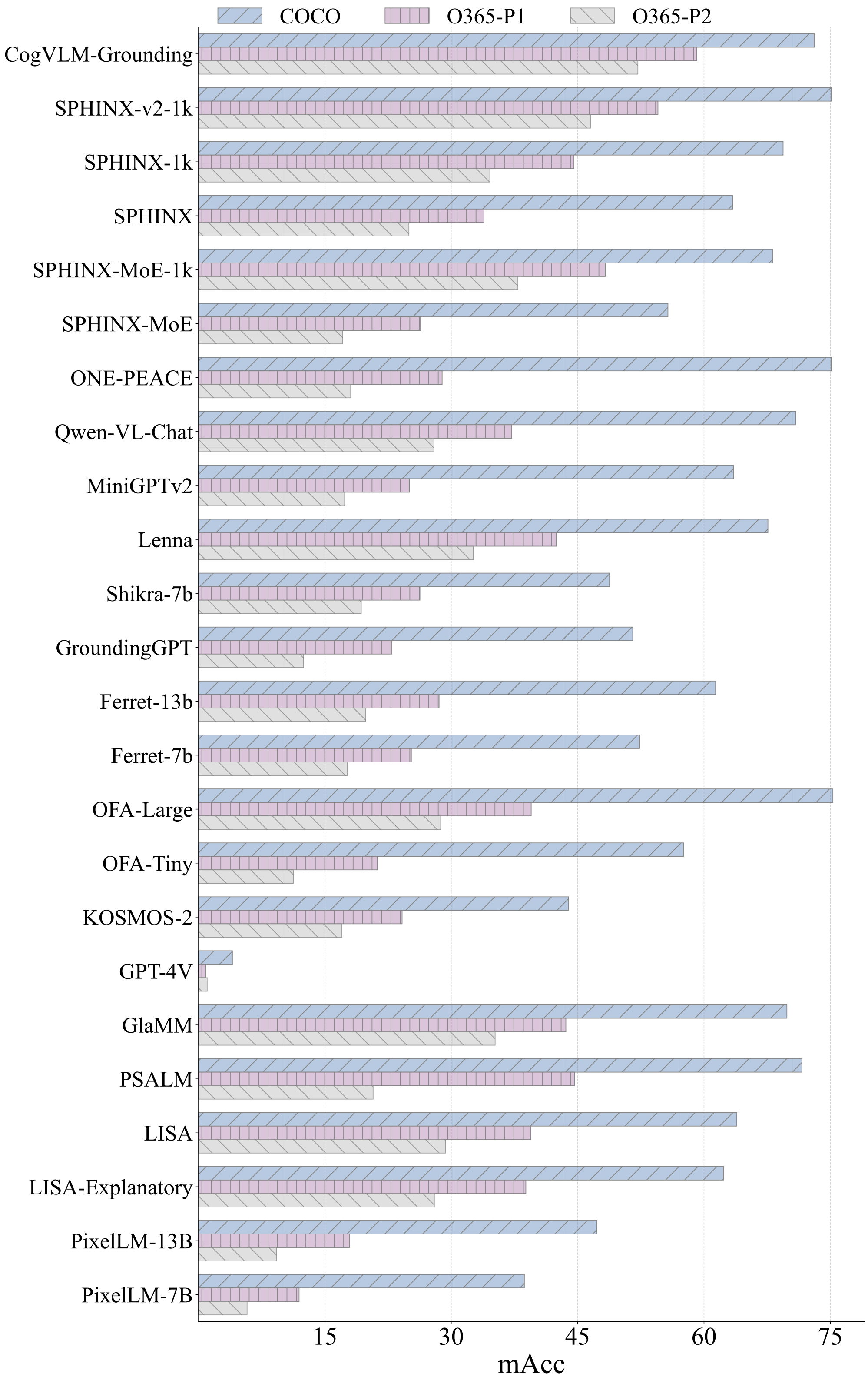


Figure 12: Evaluation of 24 models on various data sources, with mAcc acting as the metric.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Refer to Section 1.
   (b) Did you describe the limitations of your work? [Yes] Refer to Section D.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Refer to Section D.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] No theoretical results in this work.
   (b) Did you include complete proofs of all theoretical results? [N/A] No theoretical results in this work.
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Refer to Sections 3 and 4.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] No training is conducted in this work.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] No training is conducted in this work.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] NVIDIA A100 (80G) GPUs are used for evaluation in this work.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Our benchmark is derived from RefCOCO, RefCOCO+, RefCOCOg and Objects365.
   (b) Did you mention the license of the assets? [Yes] Refer to Section H.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We introduce a new benchmark, Ref-L4.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] The benchmark is labeled and reviewed by the authors of this work.