Datasheet of the Ref-L4 Benchmark

1 Motivation

1.1 For what purpose was the dataset created?

Existing REC benchmarks such as RefCOCO, RefCOCO+, and RefCOCOg suffer from several limitations such as high labeling error rates, limited vocabulary size and brief referring expressions. The introduced Ref-L4 benchmark is proposed 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 models using various evaluation protocols, including accuracy, scale-aware evaluation, and category-wise evaluation.

1.2 Who created this dataset and on behalf of which entity?

This dataset was developed by Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S.-H. Gary Chan, and Hongyang Zhang. They conducted this work on behalf of the Hong Kong University of Science and Technology, Microsoft Research Asia, the University of Sydney, and University of Waterloo.

2 Composition

2.1 How many instances are there in total?

Our final Ref-L4 benchmark encompasses 9,735 images with 45,341 referring expressions, each accurately describing one of the 18,653 unique instances.

2.2 What data does each instance consist of?

Each instance is a referring expression that describes a unique target within an image.

2.3 Is there a label or target associated with each instance?

Yes, each instance is associated with a bounding box and a detailed referring expression.

2.4 Are relationships between individual instances made explicit?

N/A. Each instance is independent.

2.5 Are there recommended data splits?

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

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

2.6 Are there any errors, sources of noise, or redundancies in the dataset?

The annotations underwent manual review to reduce errors, especially those caused by hallucinations in descriptions generated by GPT-4V.

2.7 Is the dataset self-contained, or does it link to or rely on external resources?

The dataset is self-contained.

2.8 Does the dataset contain data that might be considered confidential?

No.

2.9 Does the dataset contain data that might be offensive, insulting, threatening, or might otherwise cause anxiety?

No.

2.10 Does the dataset relate to people?

Our benchmark is constructed using four publicly available datasets: RefCOCO, RefCOCO+, RefCOCOg, and Objects365. These datasets feature images that include the "human" and "person" categories. However, Ref-L4 is not specifically designed for research focused on human-related subjects.

2.11 Does the dataset identify any subpopulations?

No.

2.12 Is it possible to identify individuals from the dataset?

No.

2.13 Does the dataset contain data that might be considered sensitive in any way?

No.

3 Collection Process

3.1 How was the data associated with each instance acquired?

Our benchmark is constructed using four publicly available datasets: RefCOCO, RefCOCO+, RefCOCOg, and Objects365.

3.2 If the dataset is a sample from a larger set, what was the sampling strategy?

For the RefCOCO 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 objects. 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.

3.3 Who was involved in the data collection process and how were they compensated?

All authors of this paper were involved in both data collection and the manual review processes.

3.4 Over what timeframe was the data collected?

The Ref-L4 benchmark was collected over a period of eight months, spanning from August 2023 to March 2024.

3.5 Were any ethical review processes conducted?

Yes.

3.6 Does the dataset relate to people?

N/A.

4 Preprocessing/cleaning/labeling

4.1 Was any preprocessing/cleaning/labeling of the data done?

Yes. 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. We manually review all referring expressions generated by GPT-4V to correct any hallucination issues. We ensure that each expression uniquely describes the instance and is factual, accurate, and harmless.

4.2 Is the software used to preprocess/clean/label the instances available?

Yes, we used GPT-4V, which is publicly available.

5 Uses

5.1 Has the dataset been used for any tasks already?

In our study, we used our Ref-L4 benchmark to evaluate 24 REC models.

5.2 What (other) tasks could the dataset be used for?

The dataset can be used for evaluating any models, and particularly the large multimodal models that are capable of handling the REC task.

6 Distribution

6.1 Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?

Yes. The Ref-L4 benchmark is available for download from the Huggingface platform at https://huggingface.co/datasets/JierunChen/Ref-L4. The DOI is: 10.57967/hf/2388.

6.2 When will the dataset be distributed?

The first version of the dataset is scheduled for release in June 2024.

6.3 Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset is licensed under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.

6.4 Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

7 Maintenance

7.1 Who is supporting/hosting/maintaining the dataset?

The responsibility for maintaining the dataset lies with the authors.

7.2 How can the owner/curator/manager of the dataset be contacted?

The owner of the dataset can be reached at jcheneh@cse.ust.hk and fawe@microsoft.com.

7.3 Will the dataset be updated?

The authors of Ref-L4 are dedicated to the ongoing maintenance and preservation of this valuable dataset. Recognizing its importance for advancing research, we plan to release future updates and expansions as the dataset is utilized in subsequent studies. Our maintenance strategy includes vigilant monitoring and prompt resolution of issues identified by the broader research community following its release. Additionally, we aim to incorporate feedback and contributions from users to ensure the dataset remains relevant and continues to meet the evolving needs of the academic community.

7.4 Will older versions of the dataset continue to be supported/hosted/maintained?

Yes.

7.5 If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Yes, the dataset will be released under a Creative Commons (CC) license, allowing others to use, reproduce, and build upon the data as long as they comply with the terms of the license. This ensures that researchers and developers can freely contribute to and enhance the dataset, fostering a collaborative environment for further advancements.

8 Reproducibility of the baseline score

We evaluated 24 open-source models. The evaluation code is provided in https://github.com/JierunChen/Ref-L4.

9 Reading and using the dataset

Instructions for accessing and using the dataset are available in https://huggingface.co/datasets/JierunChen/Ref-L4.

10 Data Format

Each entry in the Ref-L4 dataset includes an image, a bounding box, and a comprehensive referring expression annotation. These annotations are stored in parquet format. Table 1 demonstrates the key-value structure of the annotation.

Table 1: Key-value structure of the annotation.

Key	Value
id	1
caption	"Within the central picture frame ..."
bbox	$[x, y, w, h]$
bbox_area	10,492.60
bbox_id	"o365_527361"
ori_category_id	"o365_64"
image_id	"o365_922765"
height	741
width	1,024
file_name	"objects365_v2_00922765.jpg"
is_rewrite	true
split	"val"

The annotation format primarily consists of the following key-value pairs:

1. **id**: A unique identifier assigned to each triplet of an image, an instance, and an expression.
2. **caption**: A comprehensive natural language description of the target instance, offering context and detailing specific attributes.
3. **bbox**: The bounding box coordinates of the target instance, represented as $[x, y, w, h]$, where where x and y denote the top-left coordinates, and w and h signify the width and height of the bounding box, respectively.
4. **bbox_area** The area of the bounding box.
5. **bbox_id**: Unique identifier for the box.
6. **ori_category_id**: Original category identifier.
7. **image_id**: Unique identifier for the image.
8. **height**: Height of the image.
9. **width**: Width of the image.
10. **file_name**: The filename of the image.
11. **is_rewrite**: Indicator if the caption is a rewritten version, false for raw caption and true for rewritten.
12. **split**: Benchmark split ("val" or "test").