OmniLabel: A Challenging Benchmark for Language-Based Object Detection Supplemental Material

Samuel Schulter[1,†] Vijay Kumar B G[1] Yumin Suh[1] Konstantinos M. Dafnis[2,*]
Zhixing Zhang[2,*] Shiyu Zhao[2,*] Dimitris Metaxas[2]

[†] project lead [*] equal technical contribution, alphabetic order

[1] NEC Laboratories America [2] Rutgers University

1. Overview of Appendix

- **Sec. 2:** Additional analysis of the OmniLabel dataset
- **Sec. 3:** Additional information on data collection
- **Sec. 4:** Code and mini-dataset
- **Sec. 5:** Visualizations of the OmniLabel dataset

2. Additional dataset analysis

We further analyze the object descriptions we collected for the OmniLabel benchmark in the following paragraphs.

Part-Of-Speech (POS) tags: In Sec. 5.2 of the main paper, we analyze object descriptions by their POS tagging. To get POS tags, we use the `spacy` toolbox [2], which categorizes each word into one of 17 UPOS tags [1], out of which we selected the 6 most relevant tags for Fig. 5 of the main paper:

- ADJ: adjective
- ADP: adposition
- DET: determiner
- NOUN: noun
- PROPN: proper noun
- VERB: verb

Fig. 1 shows word clouds for the tags NOUN, VERB and ADJ, collected from a random subset of 5K object descriptions.

Types of language understanding: To further analyze the our object descriptions, we manually tagged a random subset of 500 descriptions with what type of language understanding they require:

- **"categories"**: The description contains one or more object category names
- **"spatial relations"**: Example: "left to", "behind"
- **"attributes"**: Attribute of objects, *e.g.*, color or material
- **"(external) knowledge or reasoning"**: Knowledge beyond the image content
- **"functional relations"**: Describing objects by their functionality, *e.g.*: "edible item" or "areas to sit on"
- **"actions"**: Any action an object can perform, "person jumping", "parked car"
- **"numeracy"**: Descriptions that require reasoning about numbers, like counting or understanding the time

Fig. 2 shows the results of our manual tagging efforts as the percentage of description that were tagged with one of the above types. Note that one description can be tagged with multiple categories. For example, the description "A black cat jumping onto the chair on the left" would get tags for "attribute" (black), "categories" (cat, chair), "action" (jumping), and "spatial relations" (on the left).

We can see from Fig. 2 that more than 80% of object description include some category name, which is expected. Note that the number of unique nouns is not limited to a fixed label space. In fact, the validation set of OmniLabel has 4.6K unique nouns, a lot more than existing benchmarks, see Table 1 of the main paper. Besides category names, close to 40% of object descriptions require an understanding of attributes, spatial relations, and external knowledge or reasoning for correct localization of objects. And finally, understanding of functional capabilities, actions and numeracy is needed for 5-10% of the descriptions. In Sec. 5, we provide visual examples for each of the above groups.

Distribution of number of boxes per description: One aspect that differentiates our OmniLabel dataset from prior benchmarks is the number of instances (bounding boxes) that are referred to by one object description. As we can see in Fig. 3, for both RefCOCO/g/+ [5, 9] and Flickr30k [6] all descriptions refer to exactly one instance in the image. PhraseCut [8] and OmniLabel allow multiple instances per description, while OmniLabel shows a lower bias towards referring to one instance.

3. Additional information on data collection

Sec. 4 of the main paper describes our data collection process. One aspect of this process is that we start from object detection datasets with existing annotations of

(a)

(b)

(c)

Figure 1: Word clouds of nouns (a), verbs (b) and adjectives (c) collected from a subset of 5K object descriptions.

bounding boxes and corresponding semantic categories. On COCO [4], semantic annotations contain a category name along with a grouping into super-categories. For Objects-365 [7] and OpenImages [3], we manually grouped categories into super-categories.

We leverage this semantic annotation when selecting images for annotation with free-form object descriptions. Specifically, we sample pairs of images and (super-) categories for annotation that fulfill some constraints (enough instances available, see Sec. 4 of the main paper). Fig. 4 shows the distribution of object descriptions over their origin:

- Plain: Original categories of the underlying dataset
- FF-Class: Free-form descriptions based on categories
- FF-SuperClass: Free-form descriptions based on super-categories

The intuition behind sampling based on different types of categories is to collect object descriptions that go beyond using the original category names along with additional context to specify subsets of object instances. And indeed, we found that 45.3% of the "FF-Class" descriptions use the underlying category name, while only 10.8% of the "FF-SuperClass" descriptions use the super-category name and only 5.3% of the "FF-SuperClass" descriptions use any of the subclass names.

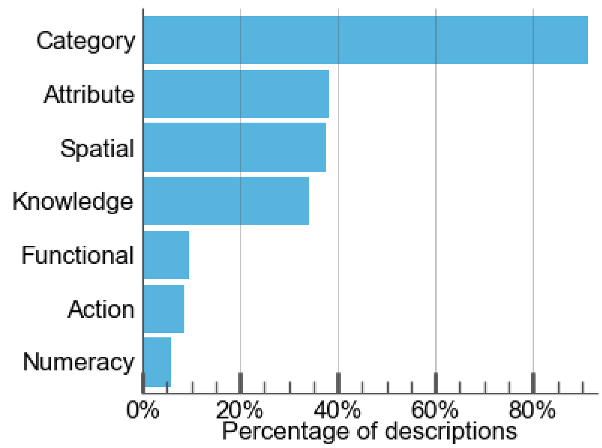


Figure 2: We manually tagged a random subset of 500 object descriptions with the types of language understanding needed to localize the referred instances correctly. The plot shows the percentage of object descriptions tagged for each type. Each description can be tagged with multiple types.

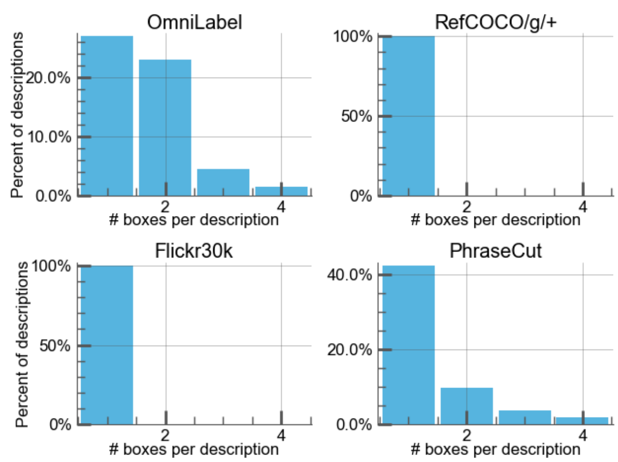


Figure 3: Distribution of object descriptions referring to different number of instances in the image.

Collection of negative object descriptions A major claim in our paper is the existence of negative descriptions. These are object descriptions that are semantically related to an image, but do not refer to any object. For any given image, we collect such negative descriptions by first ran-

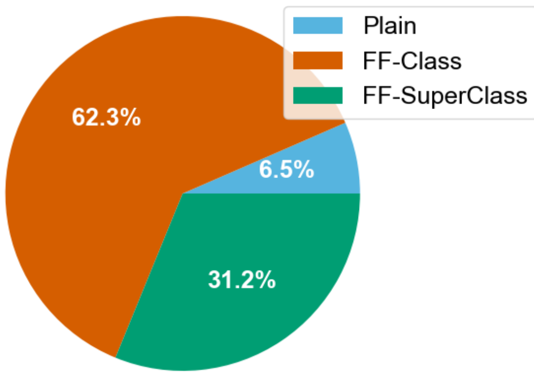


Figure 4: Pie chart showing the distribution of object descriptions grouped into plain categories and free-form descriptions collected based on standard categories (FF-Class) and super-categories (FF-SuperClass).

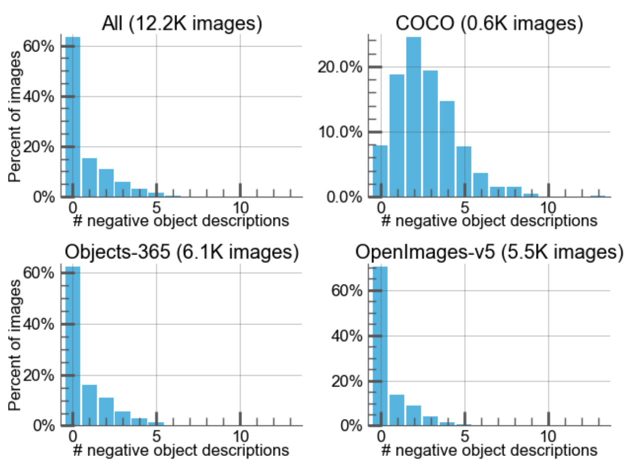


Figure 5: Distribution of images with different number of negative descriptions. The title of each sub-plot indicates the subset of images that were inspected. Images from COCO have a significantly different distribution to Objects-365 and OpenImages-v5 due to our annotation schedule, see text.

domly sampling collected positive descriptions from other images that contain the same (super-) category. Then, the randomly selected descriptions are manually verified by human annotators to not refer to any object in the image. The semantic relation to the image we obtain from the sampling process makes these negative descriptions difficult distractors. Figs. 15 and 16 in Sec. 5 show several examples.

Finally, Fig. 5 shows a distribution of the number of negatives per image, for all images of the dataset as well as for the set of images coming from the three datasets we used for annotation, COCO [4], Objects-365 [7], and OpenImages-v5 [3]. The figure shows a significantly different distribution for COCO compared to the other datasets. The absolute numbers of negatives are different given the number of images per dataset, see title of sub-plots. Still, there are two reasons for this stark difference and both relate to our annotation process. First, we collected negative descriptions only for 50% of the images in Objects-365 and OpenImages-v5[1]. Second, we found that the verification rate of negative descriptions (see Sec. 4 of the main paper) is clearly higher for COCO (around 45%) compared to Objects-365 (around 25%) and OpenImages-v5 (around 16%). We suspect the number of underlying object categories to cause this difference in the verification rates, but this aspect needs further investigation.

Nevertheless, the total number of negative descriptions in OmniLabel is currently around 10K, sufficient to make a clearly noticeable impact in the evaluation of models. This can be seen from our evaluation in Table 3 of the main paper, specifically when looking at the difference between AP-descr and AP-descr-pos. The difference between these metrics is that AP-descr-pos does not evaluate on negative descriptions. Given that we observe significantly higher numbers for AP-descr-pos, particularly for COCO images, we can safely conclude that negative descriptions pose a significant challenge to current language-based models.

Annotation interface: We provide screenshots of the annotation interface for the three tasks we rely on human annotators, see also Fig. 4 in the main paper:

(a) "collect object descriptions" (Fig. 7)
(b) "Verification of descriptions" (Fig. 8)
(c) "Collection of negative descriptions" (Fig. 9).

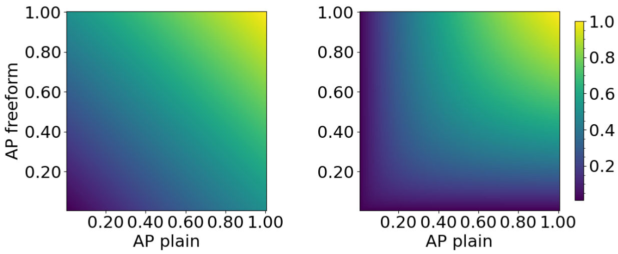
4. Code and Dataset

Along with the dataset, we built a Python-based toolkit to visualize samples from the dataset and to evaluate prediction results. The toolbox is publicly released at `https://github.com/samschulter/omnilabeltools` and includes a Jupyter notebook `omnilabel_demo.ipynb` demonstrating the use of the library. The last cell in the notebook runs the evaluation with dummy predictions. The final metric, as described in Sec. 3.2 of the main paper, is the harmonic mean between AP for plain and freeform-text object descriptions. Fig. 6 illustrates the impact of using the harmonic mean over the arithmetic mean.

5. Examples of Dataset Samples

Finally, we visualize some examples of our datasets. First, Figs. 10 to 14 showcase interesting positive examples

[1]This might change in the future when we collect more data



(a) (b)

Figure 6: Difference between (a) arithmetic and (b) harmonic mean when averaging two values. The preferred choice in our metric to average AP of plain and freeform-text object descriptions is the harmonic mean, because it encourages good performance on both types of object descriptions. This is apparent from the low values in both the upper left and lower right corners in (b).

that highlight the different types of required language understanding as described above in Sec. 2. Second, Figs. 15 and 16 show difficult negative object descriptions that are related to the image content but do not actually refer to any object. These negative descriptions pose a significant challenge to current language-based detection models. See the corresponding captions for more details.

References

[1] Universal Dependencies contributors. Universal POS tags. https://universaldependencies.org/u/pos/. 1

[2] Explosion. spaCy. https://spacy.io. 1

[3] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2, 3

[4] 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 *ECCV*, 2014. 2, 3

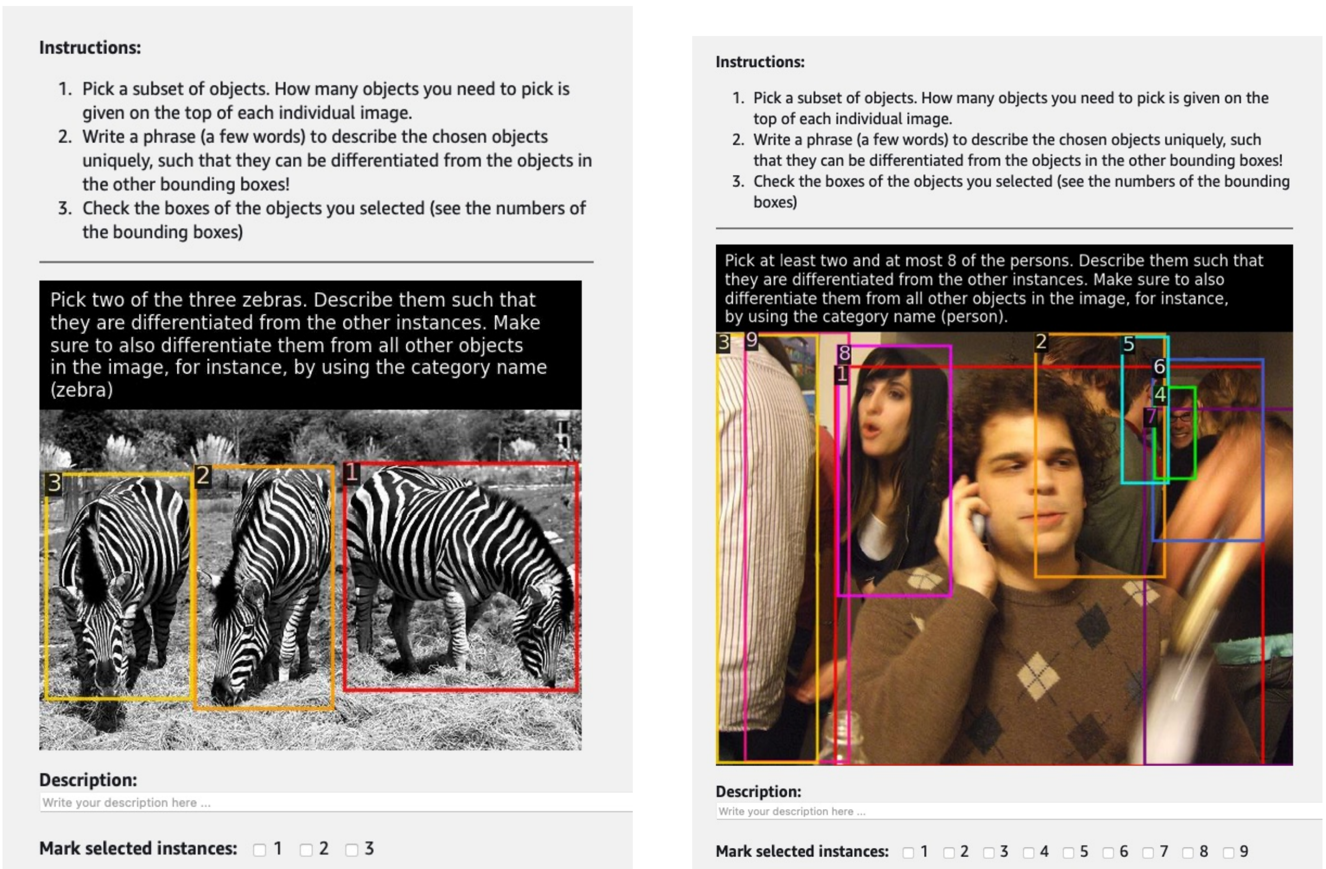
[5] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1

[6] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017. 1

[7] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Jing Li, Xiangyu Zhang, and Jian Sun. Objects365: A Large-scale, High-quality Dataset for Object Detection. In *ICCV*, 2019. 2, 3

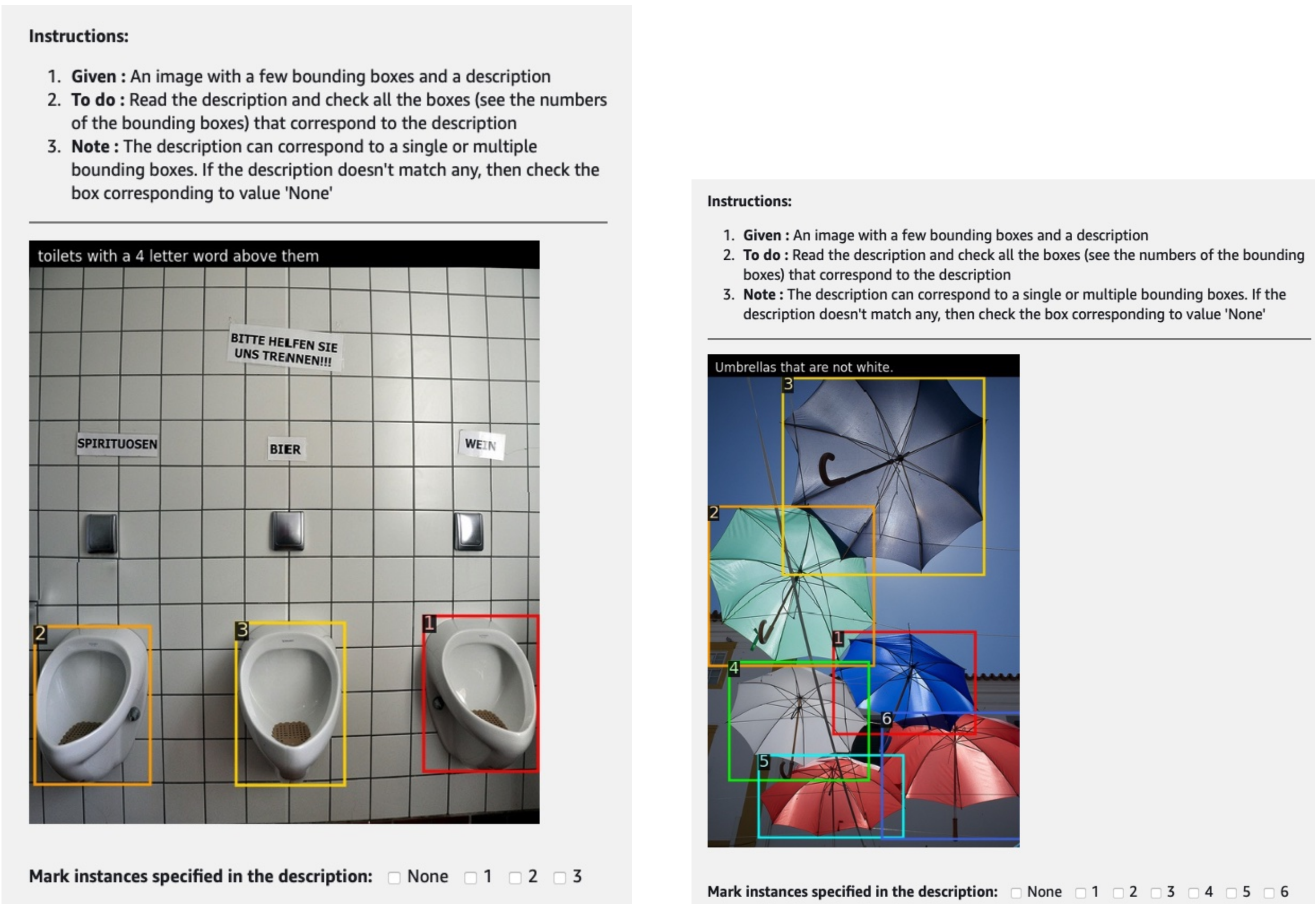
[8] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-based Image Segmentation in the Wild. In *CVPR*, 2020. 1

[9] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *ECCV*, 2016. 1



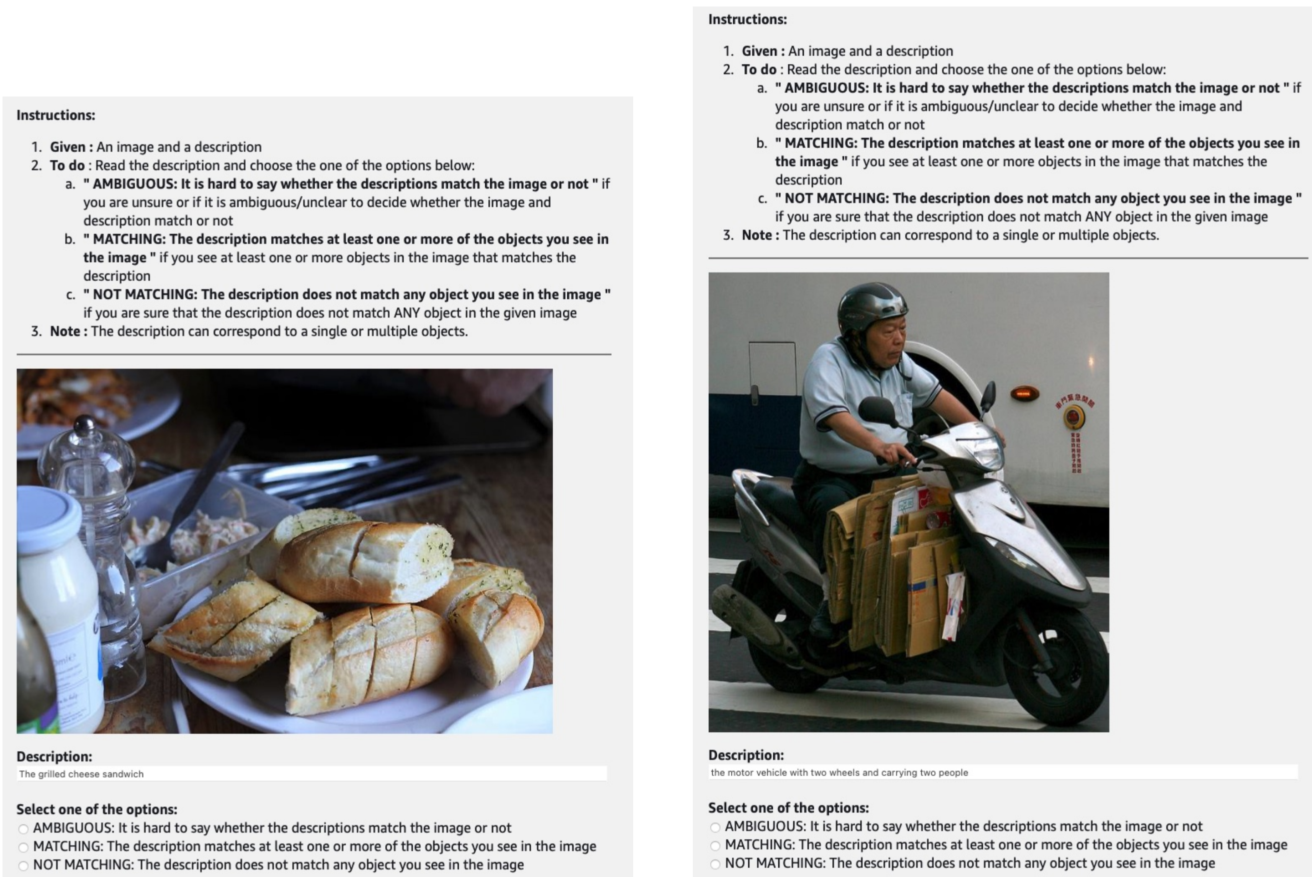
(a) (b)

Figure 7: Two examples of our annotation interface to collect object descriptions. Annotators pick a subset of the bounding boxes by clicking the corresponding checkboxes and write a freeform text description. Note that the selection has some constraints, as described in Sec. 4 of the main paper.



(a) (b)

Figure 8: Two examples of our annotation interface to verify collected object descriptions. Annotators are given the image and and a description and need to pick the matching bounding boxes by clicking the corresponding checkboxes.

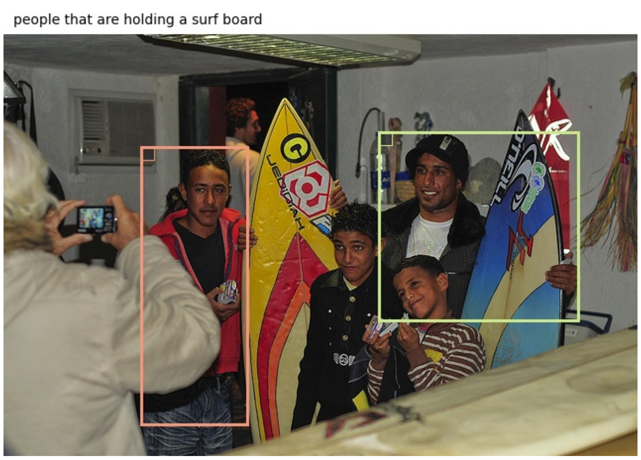


(a) (b)

Figure 9: Two examples of our annotation interface to verify *negative* object descriptions. Annotators are given an image and and a description and are asked if the object refers to any object in the image or not.

(a)

(b)



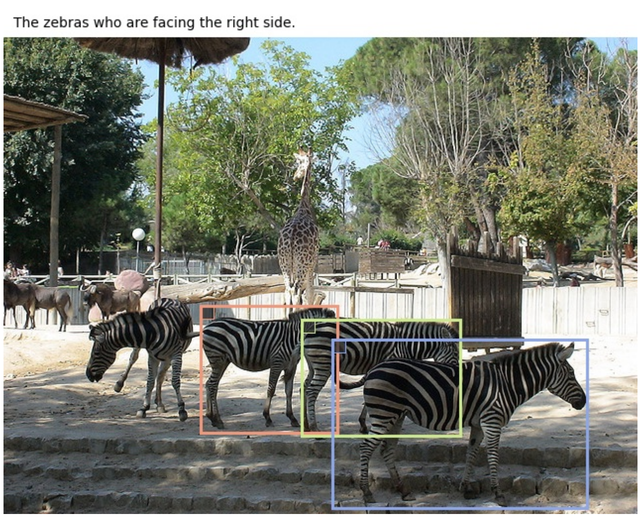
(c)

(d)

Figure 10: Examples of **positive object descriptions** requiring different types of language understanding (we only highlight a subset): categories ("adults", "benches", "woman", "skirts", "people", "surfboard") and actions ("sitting", "wearing", "holding").

(a)

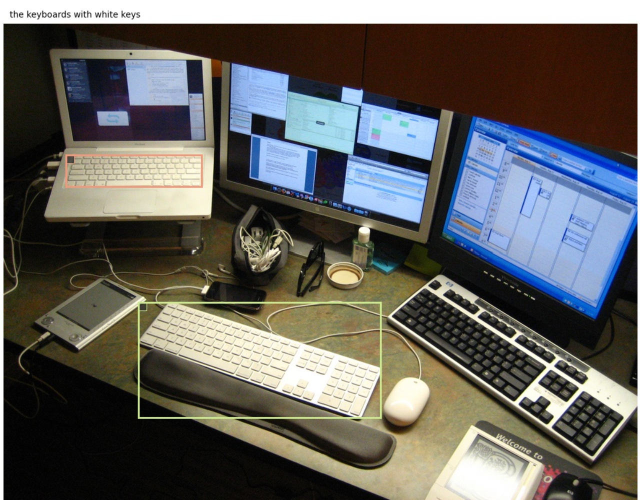
(b)



(c)

(d)

Figure 11: Examples of **positive object descriptions** requiring different types of language understanding (we only highlight a subset): spatial ("second to bottom", "closer to", "right side") and functional relations ("to drink from").



(a)

(b)

(c)

(d)

Figure 12: Examples of **positive object descriptions** requiring different types of language understanding (we only highlight a subset): attributes (“white”, “dark in color”, “green”) and numeracy (“one thirty-two”).

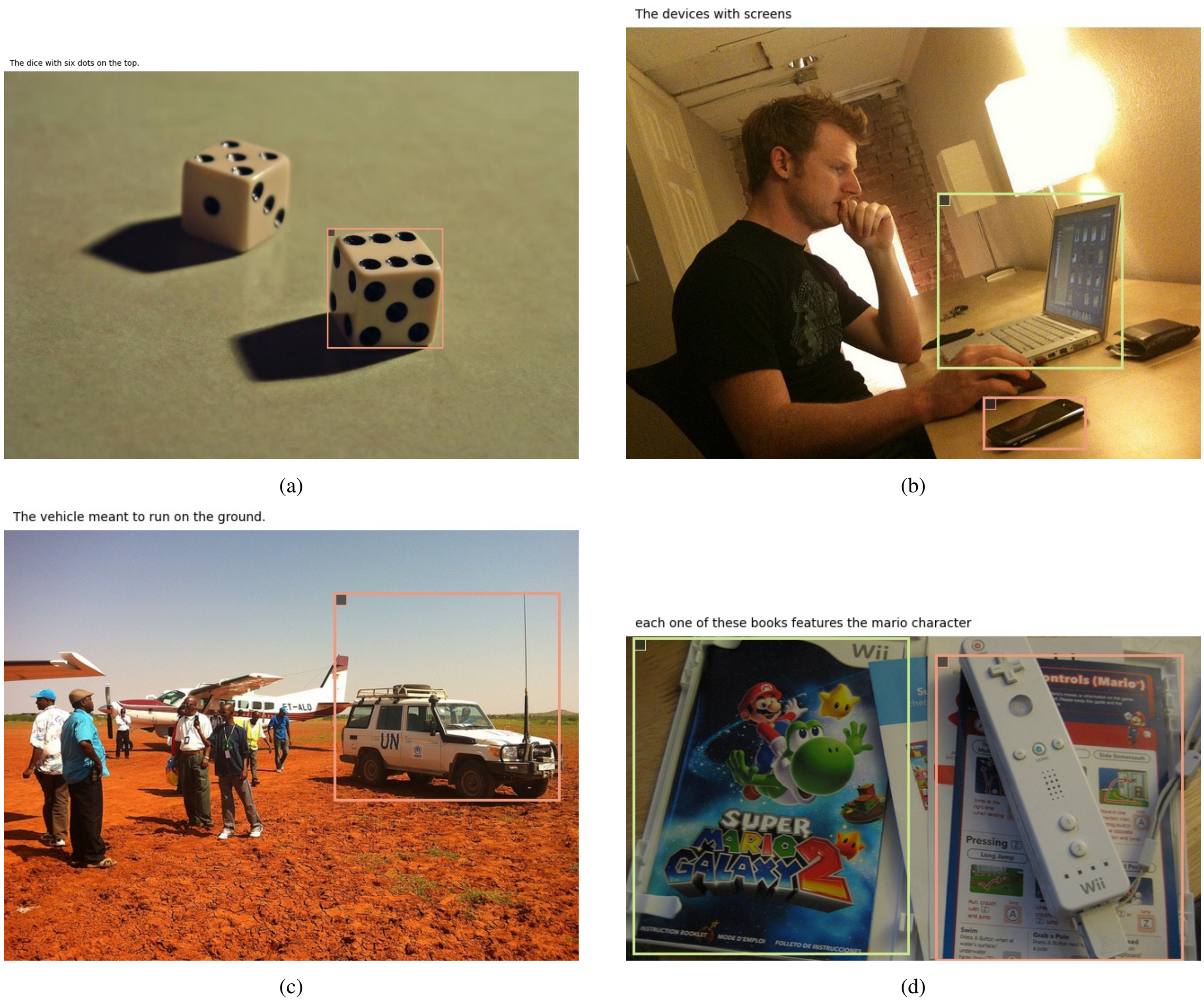
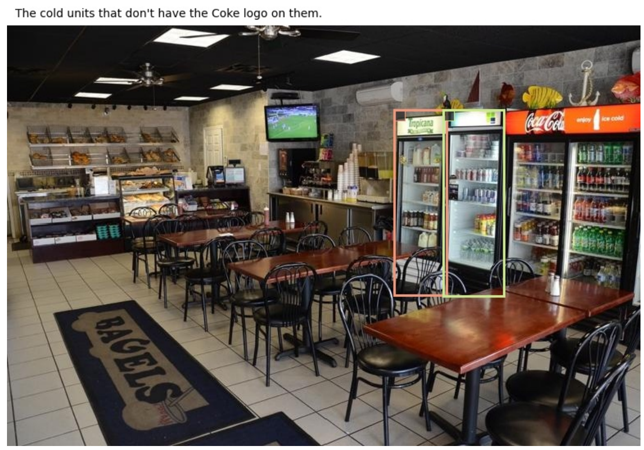


Figure 13: Examples of **positive object descriptions** requiring different types of language understanding (we only highlight a subset): numeracy (“six dots”) and external knowledge or reasoning (“devices with screens”, “meant to run on the ground”, “mario character”).

(a)



(b)

(c)

(d)

Figure 14: Examples of **positive object descriptions** requiring different types of language understanding (we only highlight a subset): external knowledge or reasoning (“container with alcohol”, “Coke logo”, “HSBC sign”, “numbered buttons”).

5 negative free-form text descriptions:
- "the flower vase"
- "The white teddy bear with the red tag on his ear."
- "The white teddy bear that is near the foot of the person."
- "an educational item that can be read and features red persons on the cover"
- "The brown wooden base with the swirls on it."

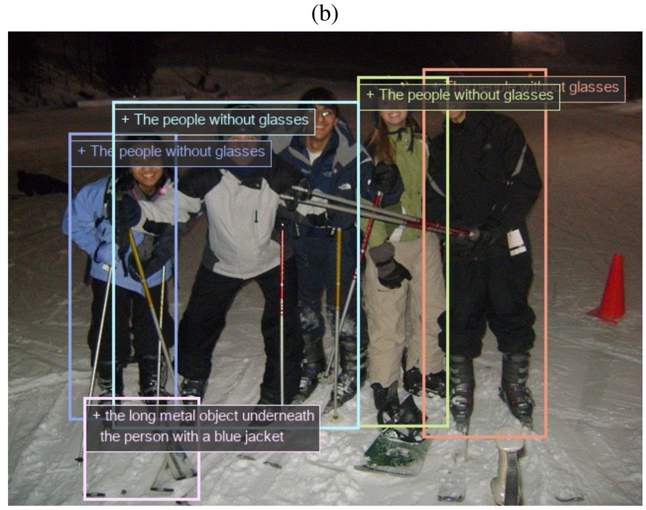
(a)

1 negative free-form text descriptions:
- "The stuffed animals that are green."

(b)

1 negative free-form text descriptions:
- "The carrot that is touching the fish."

(c)



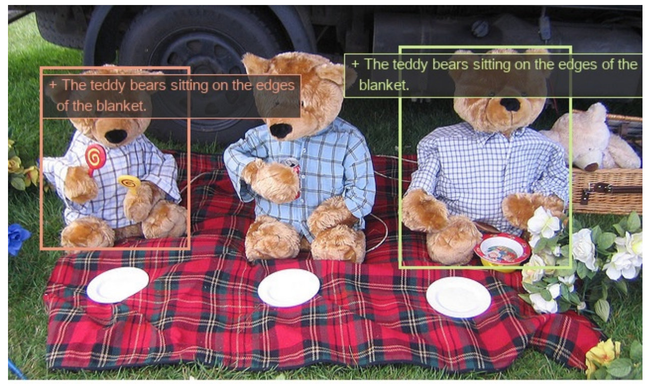
8 negative free-form text descriptions:
- "Ballplayers wearing shirts with contrasting sleeve color starting at shoulders."
- "Person with food in front of them"
- "A person wearing torn clothing."
- "All the people holding umbrellas"
- "a leather piece of equipment that allows you to catch things"
- "Sport item you hit balls with"
- "Item you stand on with wheels"
- "Item you put on hand to catch ball"

(d)

Figure 15: Examples of difficult **negative object descriptions**, which are listed below the respective images. Note that for positive descriptions, we only show the freeform-text descriptions and omit the plain categories to avoid cuttered visualizations in the image.

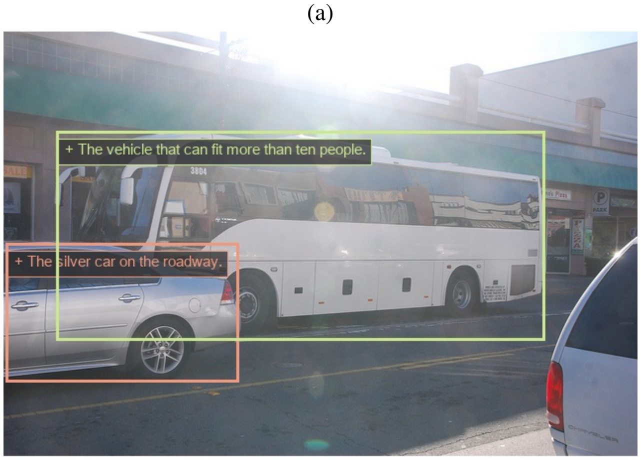
2 negative free-form text descriptions:
- "The racket the woman is holding."
- "The tennis racket being held by the man in red."

(a)



3 negative free-form text descriptions:
- "The teddy bear wearing a green hat."
- "a side profile of a teddy bear looking to the right"
- "The teddy bear that is furthest right, and sitting on another teddy bear."

(b)



6 negative free-form text descriptions:
- "The cars that are behind the red car."
- "Car next to a tree"
- "people watching the dog jump"
- "a green automobile for multiple passengers"
- "The vehicle with a 22 on the front of it."
- "These vehicles run on tracks rather than roads."

(c)

2 negative free-form text descriptions:
- "The people that are sitting down inside."
- "The person kneeling on the ground."

(d)

Figure 16: Examples of difficult **negative object descriptions**, which are listed below the respective images. Note that for positive descriptions, we only show the freeform-text descriptions and omit the plain categories to avoid cuttered visualizations in the image.