TRICD: Testing Robust Image Understanding Through Contextual Phrase Detection

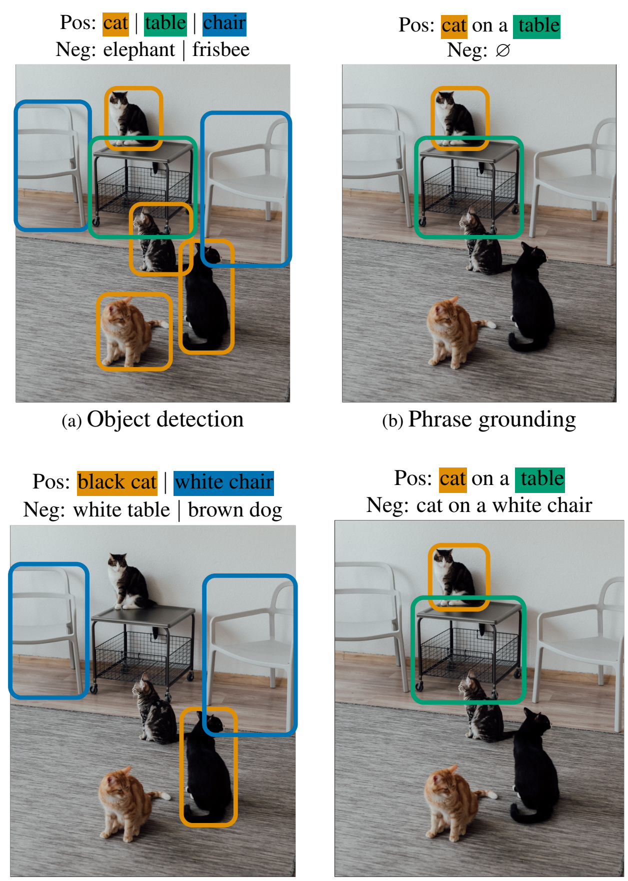
Aishwarya Kamath*†, Sara Price*†, Jonas Pfeiffer†, Yann LeCun†, Nicolas Carion*†

†New York University

{aish,sbp354,yann}@nyu.edu

Abstract

Most traditional benchmarks for computer vision focus on tasks that use a fixed set of labels that are known a priori. On the other hand, tasks like phrase grounding and referring expression comprehension make it possible to probe the model through natural language, which allows us to gain a more extensive understanding of the model's visual understanding capabilities. However, unlike object detection, these free-form text-conditioned box prediction tasks all operate under the assumption that the text corresponds to objects that are necessarily present in the image. We show that results on such benchmarks tend to overestimate the capabilities of models significantly given that models do not necessarily need to understand the context, but merely localize the named entities. In this work we aim to highlight this blind spot in model evaluation by proposing a novel task: Contextual Phrase Detection (CPD). To evaluate it, we release a human annotated evaluation dataset called TRICD[1]. It consists of instances of two image-text pairs with bounding boxes for each of the phrases present in the image. The pairs are contextually related, but partially contradictory; i.e. while the images and texts are semantically similar, each sentence is only depicted in one of the images, but not the other. Models must predict the relevant bounding boxes for the phrases in an image if and only if it is in accordance with the context defined by the full sentence. We benchmark the performance of several state of the art multi-modal models on this task in terms of average precision (AP).
Website : https://ashkamath.github.io/TRICD/



(a) Object detection (b) Phrase grounding (c) Phrase detection (d) Contextual Phrase Detection

Figure 1. **Contextual Phrase Detection** (1d) extends previous related tasks: Like object detection (1a), it evaluates both positives and negatives; like phrase detection (1c), it has an open vocabulary; and like phrase grounding (1b), the context surrounding the phrases is important.

1. Introduction

Understanding visual scenes is a fundamental objective in the field of computer vision. Over the years, several proxy tasks have been proposed to quantify how well models fully grasp the contents of an image from image level tasks such as classification [13, 31] to dense prediction tasks such as object detection [19, 33, 39, 59], segmentation [11, 12, 17, 29] and depth prediction [61]. These benchmarks provide a valuable north-star for researchers in the quest to build better visual understanding systems. A limita-

[1]Testing Robust Image understanding through Contextual Phrase Detection (pronounced "tricked")
* indicates equal contribution

tion of these traditional computer vision benchmarks, however, is that they typically restrict their label sets to a fixed vocabulary of concepts known *a priori*. This inherently creates blindspots and biases in the set of capabilities that models can obtain and be evaluated on.

To relax this rigid formulation, one possibility is to design benchmarks that *leverage natural language* to probe a model's understanding of a given image in a more nuanced manner. One of the earliest such tasks is image captioning [65], followed by many others such as Visual Question Answering (VQA) [4, 25, 26, 62], Visual Commonsense Reasoning (VCR) [74], Visual Entailment (VE) [69], inter alia. We are particularly interested in tasks that probe model's fine-grained localization capabilities such as phrase grounding [54] and referring expression comprehension (REC) [1, 28]. While they form a natural extension of classical object detection, these tasks assume that the objects of interest are visible in the image thus boiling them down to just localization, and not true object detection.

In this paper, we propose a bridge between these two types of tasks that we term *Contextual Phrase Detection (CPD)*. In CPD, models are provided with one or more phrases that may be part of a larger textual context; the model must detect all instances of each phrase if and only if they are in accordance with the context defined by the full sentence. For example, given a sentence "cat on a table", we require the model to predict boxes for any *cat* and *table* where there is a *cat on the table*, and for no other object (including other cats or tables that may exist in the image; see Figure 1d). Crucially, and differently from REC and phrase grounding, *we do not assume a priori that all phrases are groundable*. Relaxing this assumption tests the model's ability to refrain from predicting boxes if no object satisfies the constraints specified by the whole sentence. This can be seen as a true generalization of the object detection task, since proficiency in both localization (where the objects are) and classification (is the mentioned object present?) are required to solve the task. CPD opens the door to evaluating models' detection capabilities in a truly flexible way: instead of being constrained by the vocabulary, we can now *benchmark the detection of anything that can be described in free form text*.

To support the evaluation of this novel task, we release TRICD, a human-annotated evaluation dataset of 2672 image-text pairs having 1101 unique phrases associated with a total of 6058 bounding boxes. An important requirement for accurately measuring the model's ability to determine if an object specified by a phrase is present in the image is having explicit negative certificates for a phrase given an image. We extend the previous efforts towards open-ended detection [53] with this added constraint. It is intractable to obtain negative certificates for all the phrases in all the images, hence we follow the trend in large-vocabulary detection benchmarks [19] and take a *federated* approach: for each positive phrase, we carefully select a related "distractor" image in which the target phrase does not occur. The main hurdle lies in the procurement and verification of such negative instances, especially those that can truly test a model's discriminative abilities. We emphasize finding challenging negatives by ensuring that the distractor image shares some core traits with the positive one (for example, having a similar scene).

(a) Is a person rowing in the river? (b) Is there a baseball bat?

Figure 2. Questions where SOTA VQA models answer "yes".

Our experiments on TRICD demonstrate that state-of-the-art (SOTA) multimodal systems that achieve impressive performance on numerous downstream tasks (e.g. REC [16, 27, 37, 66, 75], VQA [2, 66, 67], and phrase grounding [16, 27, 37, 75]), still demonstrate a lack of robustness when presented with more confusing or ambiguous image-text pairs. We find that models often misidentify objects when they appear in surprising contexts or hallucinate non-existent objects depending on their surroundings. This finding is reminiscent of hallucination phenomena in image captioning systems [57]. For example when asked "Is there a person rowing a boat in the river?" about Fig 2a, and "Is there a baseball bat?" about Fig 2b, SoTA VQA models like FIBER, OFA and Flamingo-3B all answer "yes". CPD requires predicting bounding boxes, which allows a more fine-grained understanding of reasoning processes and failure modes of VL models.

We show that there is a large performance gap ($\sim$10 points) between the evaluated models' performance on TRICD compared to benchmarks like GQA [25] and Flickr30k [54] when compared in terms of F1-score on binary questions and phrase grounding recall@1 respectively, indicating that our dataset is *challenging*. On the CPD task, the best model achieves 21.5 AP on TRICD. We examine failure cases and find that there is substantial room for improvement in SoTA models' abilities to understand contextual cues. We hope that TRICD serves to better measure progress on building visual understanding models having fine-grained spatial and relational understanding.

2. Related Datasets and Benchmarks

The datasets available to us largely determine the capabilities with which we can equip our models and provide a means for measuring progress. Seminal works introducing datasets like Imagenet [13], COCO [39] and Flickr30k [72] drove forward research along several axes such as large scale image classification, classification and localization of objects in images, prediction of segmentation masks, and image-text retrieval. In this section we describe some datasets and associated evaluation benchmarks that are the most related to our goals and those which as well as those which informed several of our design choices.

Visual Grounding. The task of visual grounding consists of predicting bounding boxes corresponding to a plain text caption. There are two main variants of this task: **Phrase grounding**, which involves predicting boxes for each noun-phrase of a caption and **Referring Expression Comprehension** (REC), which involves predicting a single bounding box corresponding to the full sentence. Phrase grounding is evaluated on the Flickr30k Entities dataset [54], which consists of 30k images annotated with 5 captions having bounding boxes for each noun phrase. There are several datasets for REC, such as RefCOCO, RefCOCO+ [28] and RefCOCOg [47]. More recently, Ref-Adv [1], an adversarial split of the RefCOCOg dataset was introduced, probing for the model's sensitivity to word order. Here we stress the fact that for visual grounding, it is assumed that the phrases being queried do occur in the image. On the contrary, our proposed task (CPD) is harder since it involves a preliminary step of checking whether the phrase appears in the image. Current state-of-the-art models have come close to just 10% error rates on Flickr30k Entities (see Table 5), and perform similarly well on REC datasets. It has not been explored whether these excellent grounding abilities transfer to good detection performance in generalized CPD.

LVIS. With more than a thousand categories, LVIS [19] is a detection and segmentation dataset that enables training and evaluation of models on an order of magnitude more concepts than previously possible. Due to the Zipfian distribution of categories in natural images, annotating data and evaluating models on a large scale vocabulary comes with inherent challenges. To address those, [19] introduced the concept of a *federated* dataset, where each category is annotated only on a subset of images. Our work can be seen as the natural extension of this effort towards evaluating detection performance on an ever-growing vocabulary. By replacing categories with contextual phrases, we seek to evaluate detection of anything described in plain text.

Phrase Detection. Recently, several works [53, 77] proposed an evaluation task closely related to ours in which given a query phrase, the goal is to identify every image region associated with that phrase within a given dataset of test images. Contrary to our work, they do *not* consider context, but only the phrases themselves, thereby limiting the aspects of visual reasoning that can be evaluated (eg. complex relations as in Fig. 1). More importantly, the evaluations in [53, 77] rely on existing datasets such as Visual Genome [30] which provide regions annotated with short captions and Flickr30k Entities [54], which extracts phrases from captions. These datasets do not provide an explicit negative certificate for phrases. Rather, they rely on an *implicit* signal: if a phrase is not explicitly described in an image — up to synonym replacement — then it is considered a negative. However, we argue that obtaining reliable negatives this way is unsatisfactory. Without additional annotations, it is often impossible to determine whether a phrase is a true negative. If we consider an image where the phrase "cat" occurs, since it is under specified, one cannot ascertain whether the phrase "black cat" is a positive or a negative for this image. Further, since neither VG regions nor Flickr30k captions (which tend to focus on the most salient objects) are exhaustive, it is trivial to think of an image-text pairing where an object is in the background of a scene but not mentioned in the caption. This again defeats efforts to certify if a phrase is indeed a negative. By contrast, in our dataset we focus on collecting *explicit* negative certificates, thereby allowing robust detection evaluation.

Winoground [64] evaluates model's visio-linguistic understanding by asking them to match the correct pairs given two images and two captions where there are 800 correct and 800 incorrect pairings. The difficulty of this task lies in the fact that the two captions use the same set of words but differ in word order, and most SOTA models currently perform barely better than chance on this dataset. We extend these annotations by turning them into a CPD dataset.

Attribute Prediction. Closely related to our task, attribute prediction probes models' understanding of object properties beyond categories. Several datasets have been proposed [18, 38, 42, 43, 46, 50, 68]. The VAW dataset [51] is one of the largest, with 72k images annotated with 620 unique attributes for over 260k object instances. More recently, the LSA dataset [52] combines images from more sources such as Flickr30k [72], COCO [39] and OpenImages [33] to create a larger visual attribute detection dataset.

Relation Prediction. In addition to attributes, models ought to be able to recognize relationships between objects. Several datasets evaluate this ability, either with grounding [9, 20, 34, 46, 55, 58] or without [8, 71, 79]. The SVO-Probes dataset [22] evaluates models' understanding of relationships decomposed as Subject, Verb, and Object triplets. It carries out counterfactual testing by crafting pairs of images where only one element of each triplet varies. Performance of SoTA models indicates that verb understanding is the most challenging. We draw inspiration from this counterfactual design to create the relation split of our dataset.

3. Dataset design

3.1. Task definition

A CPD dataset of size N is defined as a set of pairs $\{(I_i, C_i)\}_{i=0}^{N}$ where I_i is an image and C_i is a text caption. A set of non-overlapping phrases $\mathcal{P}_i = \{P_{i,j}\}_{j=0}^{M_i}$ (where M_i is the number of phrases in C_i) is associated with each caption. These phrases are known *a priori*, and each phrase corresponds to a set of words in the caption that refer to a particular object (e.g. "a brown cat"). We note that it is not necessary for all noun-phrases to be represented in $\mathcal{P}_i$, and in particular it is natural to omit non-visual or non-groundable phrases (e.g "a sunny day"). Each phrase induces its own *contextual detection task*, where the goal is to detect all the instances associated with the phrase while satisfying the constraints imposed by the rest of the context. The context can be seen as a *filtering operator*, as it imposes additional criteria on the set of objects to be detected. *In particular, if some aspects of the context are violated, then the set of candidates becomes empty and nothing should be detected for this particular phrase.* The output expected for the contextual detection task is a set of bounding boxes that localize the objects corresponding to the phrase, if any. If a phrase corresponds to several distinct countable objects (e.g. "several cats"), then all the corresponding objects should be detected with a bounding box.

3.2. Metrics

Following practice in object detection datasets [19, 39], we choose to rely on Average Precision (AP) as our main evaluation metric. In the following, we detail how this metric is computed in the context of CPD.

For a given (I_i, C_i) pair for which the phrases of interest $P_{i,j}$ are provided, we require models to output a set of predictions, consisting of a set of bounding boxes, along with a confidence score and the ID of the phrase that each box corresponds to. We first sort all the predicted boxes for this image by decreasing confidence, keep only the 100 most confident ones, then greedily match them to the ground truth boxes. A candidate box can be matched to a ground truth box if and only if: (1) the ground truth box hasn't been matched yet (to a higher confidence candidate box) (2) the Intersection-over-Union (IoU) between the candidate and target is higher than a threshold τ and (3) the predicted phrase ID matches the phrase ID of the target. After the matching, all unmatched targets become False Negatives (FN) and unmatched predictions are False Positives (FP).

Following this, we obtain the Precision-Recall curve over the whole dataset, and measure the area under the curve, which gives us the Average Precision. Following the COCO protocol, we compute AP at 10 different IoU thresholds τ, linearly spaced in [0.5, 0.95], and average them.

The main difference with a standard detection task is that when the task involves a fixed set of classes of interest, the metric usually involves computing a different Precision-Recall curve for each category, then averaging the resulting APs, yielding a Mean Average Precision (mAP). By contrast, in CPD each phrase and its associated context induces its own detection target. Since we usually have only one datapoint where a given phrase (taking into account its context) is positive (i.e. it has some associated ground-truth boxes) and one where it is negative (we guarantee that there is no occurrence of it in the image), computing the AP using only these two datapoints would be impractical since it would be very unstable. For this reason, we use phrase IDs only during the matching process and ignore them when computing a single PR curve for all phrases in the dataset.

4. Dataset construction

We rely on two main sources for the images:

Winoground [64]: The Winoground dataset consists of 800 images with an associated sentence. All the datapoints work in pairs, where the two sentences in the pair are semantically similar, often even consisting of the same sub-words (eg. "fire truck" and "truck fire"). The images were obtained from a commercial image bank (Getty) and licensed for research purposes. Due to the nature of the collection process, the images are particularly adequate to test understanding of a specific concept with minimal confounding factors.[2] However, the image distribution is skewed towards aesthetically pleasing images, with generally low clutter and overall clear salient objects.

COCO [39]: Additionally, we seek a more "natural" image distribution to measure performance in settings that are closer to real-world images. We opt to use images from COCO: overall, the images are more diverse in quality and content than stock pictures, and often contain cluttered scenes with no clear salient object.[3]

4.1. Annotation process

We aim to construct a dataset that is organized into pairs of visually related datapoints where image pairs share some core traits (for example, having a similar scene). Given these pairs, if (I_0, C_0) is the first image and its associated caption, and (I_1, C_1) is the related datapoint, we aim to use the caption C_0 as a *positive* target for I_0 and as a *negative* target for I_1 and vice versa.

[2] Stock pictures often come in series where the actors exchange roles, while the situation stays the same. Annotators had access to the Getty Images API, allowing precise search queries to select the related image.

[3] To avoid train set contamination, we obtained permission from the COCO committee to annotate images from the test set.

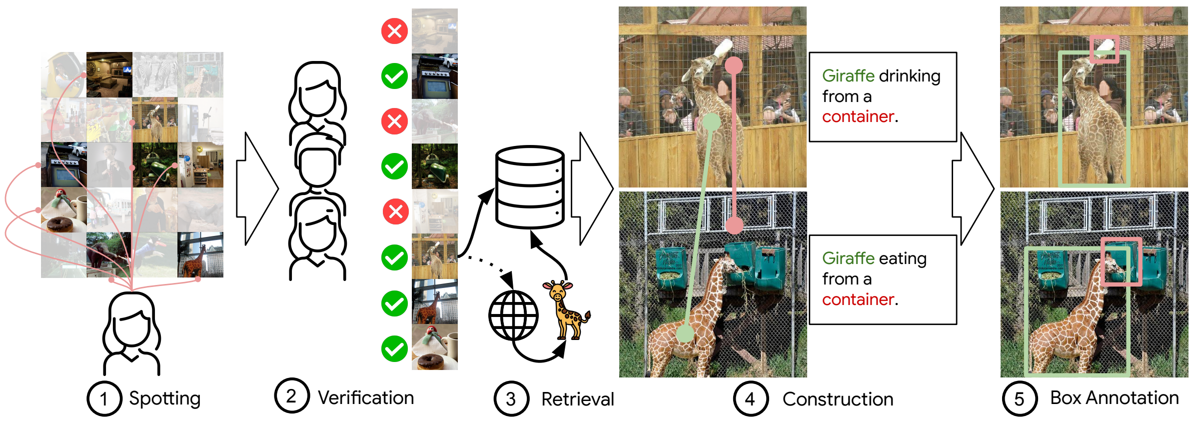


Figure 3. Annotation Process: ①: Spot images with interesting concepts; ② Verify and score spotted images; ③ Identify contextually similar images with different concepts; ④ Construct captions, extract noun phrases; ④ Annotate bounding boxes for noun phrases.

4.1.1 COCO split

We illustrate our annotation process for the COCO data in Figure 3 and describe each step below.

① **Spotting**: The annotators are presented with random images from the target set of images.[4] They are asked to list any *surprising object* or *relationship* in the image. *Objects* are defined as surprising when they are considered out-of-context in the given image, or very uncommon. *Relations* are considered surprising when (Subject, Verb, Object) triplets are rare. Annotators are encouraged to pay attention to non-salient elements. If no surprising objects or relations are found, the image is discarded. Each image is presented to at most one annotator. Approximately 5% of the images are retained in this step.

②**Verification**: For each image that was spotted in ①, a different set of annotators is asked to rate the images on a likert scale from 0 to 5. We ask the raters to filter out any ambiguous or non-groundable proposal. We ensure each proposal receives ratings by 3 annotators; we discard those with a score lower than 2.5. Overall, 50% of the candidate datapoints are discarded in this stage.

③ **Retrieval**: In the retrieval stage, given a source image I_0 and an associated query C_0, the annotators are required to find a related image I_1 and construct a query C_1, such that C_1 occurs in I_1 but not I_0 and vice versa. Crucially, we opt to entirely abstain from using any sort of *multimodal* search engine or retrieval system to select the related pictures. In doing so, we avoid inheriting potential biases or blind spots from such a system. Instead, we give the annotators access to an *image only* retrieval system, based on ConvNext [44] embeddings. This provides the annotator with the 60 images closest to the source image. Alternatively, the annotator can upload an image found by any means (eg by searching the internet), and this image will be used to retrieve the 60 closest images in the dataset.[5] For relation-based queries, we further impose the constraint that only one of the Subject, Verb, or Object differs in C_0 and C_1.

④ **Construction**: Overall, we ask annotators to craft *hard* negatives, either by finding objects and relations that would be more likely given the context, or by finding visual distractors or closely related categories. Next, we automatically extract the noun phrases from each caption using spacy.io [24]. Finally, all data points obtained undergo a last manual quality verification step, to correct spelling mistakes and ensure the validity of the queries, especially the negative pairs, which tend to be wrong in subtle ways.

⑤ **Box Annotation**: Lastly, to obtain bounding box annotations, we rely on Amazon SageMaker and Amazon Mechanical Turk (AMT). Each phrase constitutes its own task (one HIT), where we provide the workers with the image and the full sentence, along with an indication of the target noun phrase. The price per HIT is set according to the complexity of the image, and we ask three workers to annotate each image.[6] Finally, we reviewed the annotated bounding boxes using Label Studio [3], and manually improved the tightness of the bounding boxes. We do this to ensure high-quality boxes that can be used for evaluation at high IoU threshold, similar to mainstream detection datasets [19, 39].

4.1.2 Winoground split

For Winoground the process is slightly simpler given that instances are already grouped by semantically similar image and caption pairs. Therefore, we skip steps ①-③ of Figure 3. However, in ④ we perform a manual filtering step to check whether, in a given pair, both negative pairs are in-

[4]Test2017 for the test set, Val2017 for the validation set.

[5]Note however, that these uploaded images are only a means of retrieving similar images in the target dataset; the uploaded images are discarded subsequently.

[6]See the Appendix for screenshots of the annotation interface and details about the annotation worker wages.

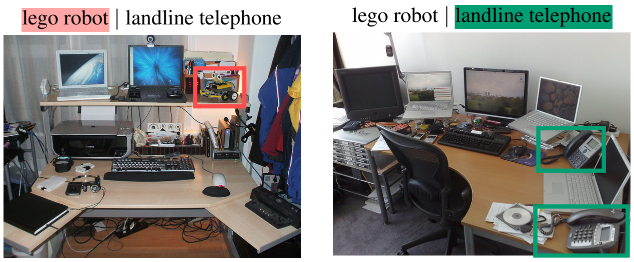


Figure 4. An example image-text pair from the COCO objects split of our validation set. The first image is a positive for the first text and negative for the second text and vice versa.

deed valid, i.e. C_1 does not appear in I_0 and C_0 does not appear in I_1. In a minority of cases, we slightly reformulate the sentences, either to make the detection target unambiguous, correct typos from the original dataset, or ensure that the negative pairs are valid. We also manually verify their correctness, and filter those that are not groundable (e.g. "a sunny day"). The ones that cannot be easily modified to fit our constraints are filtered[7]. We follow with extracting noun-phrases ④, and annotating boxes ⑤.

4.2. TRICD dataset statistics

After all these steps, we end up with 2672 image caption pairs having 1101 unique phrases with 6085 boxes. Detailed statistics of our novel TRICD dataset can be found in Table 1. We visualise the spatial distribution of the bounding boxes across the dataset in Fig. 5a and the distribution of number of boxes per phrase in Fig. 5b.

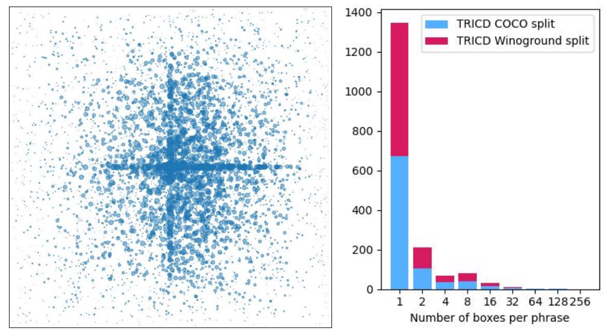
Stats	Wino	Coco Obj	Coco Rel	All
# Unique images	712	345	248	1293
# Unique phrases	706	311	196	1101
# Unique words	874	371	354	1285
# Im-cap pairs	1424	748	500	2672
# Boxes	4365	914	779	6058
Avg phrases/image	2.3	1.0	2.0	1.9
Avg boxes/phrase	2.7	2.4	1.6	2.4
Avg words/caption	9.3	1.4	5.2	6.3

Table 1. Statistics of the TRICD test set.

5. Evaluation

We aim to have a broad coverage of models for our evaluation, and choose models based on performance on standard detection benchmarks like COCO and LVIS, as well as on Phrase Grounding and Referring Expression Comprehension. For models that are primarily focused on open-vocabulary detection with an emphasis on large-scale pre-training, we use OWL-ViT [48] and DETIC [78]. For models that perform text conditioned detection and have SoTA

[7] 10% of the datapoints are filtered and 20% are edited



(a) Spatial distribution of boxes (b) Number of boxes per phrase

Figure 5. Statistics of the bounding boxes in TRICD. (a) Spatial distribution of the bounding boxes in the TRICD test set. The size of the marker represents the size of the box. (b) Distribution of box per phrase in TRICD across the two splits. The counts are quantized into bins collecting all items falling between consecutive limits on a logarithmic scale

Model	Backbones	I-T Pairs	GND Pairs	Obj Det
MDETR	RoBERTa - ENB5 [40, 63]	0	1.3M	0
GLIP-T	BERT - Swin [14, 41]	0	1.3M	600K
GLIP-L	BERT - Swin [14, 41]	0	27M	9.8M
FIBER	RoBERTa - Swin [40, 41]	4M	1.3M	600k
DETIC	CLIP - CLIP [56]	400M	0	100k
OWL-ViT	CLIP - CLIP [56]	400M	80k	2.5M
Flamingo	Chinchilla - NFNet [5, 23]	1.9B	0	0
OFA	BART-ResNet152 [21, 35]	20M	3.2M	2.98M

Table 2. Architecture of the evaluated models and pre-training data size, in Image-Text (I-T) pairs, Grounded (GND) Image-Text pairs and images from Object Detection datasets

performance on visual grounding, we use MDETR [27], GLIP [37] and FIBER [16]. We provide a brief overview of these models.[8]

MDETR is an end-to-end object detection pipeline built on DETR [6] and conditioned on free-form text. It predicts bounding boxes and which words in the input caption they correspond to. MDETR has not been trained on negative examples (e.g. through object detection data) and hence is expected to perform poorly on the negatives in our dataset.

GLIP casts object detection as a grounding task and incorporates both detection and grounding data in its training.

FIBER extends GLIP and leverages coarse-grained image-text pre-training for subsequent fine-grained image understanding by having a fused backbone architecture that integrates the image and text modalities deeper in the model compared to MDETR or GLIP.

DETIC is an open-vocabulary detector that uses CLIP [56] embeddings to encode the class names. It leverages a mixture of box-annotated data as well as image-level annota-

[8] For details please see the Appendix §G and Table 2.

Model	TRICD			
	Wino	COCO objects	COCO relations	All
Grounding models				
MDETR	10.1	3.9	20.4	10.7
GLIP-T	14.7	22.5	25.1	16.8
GLIP-L	18.1	26.9	28.6	20.1
FIBER	**19.1**	25.3	**31.6**	**21.5**
Open vocabulary detection models				
OWL-VIT	6.3	13.7	16.3	7.9
DETIC	8.7	**27.0**	19.7	11.6

Table 3. Average Precision (AP) score on subsets of TRICD

tions from ImageNet, with a weakly-supervised loss.

OWL-ViT also relies on CLIP, relying on a very large VIT [15]. During fine-tuning, it uses object detection datasets to train a localization head, using a matching loss similar to DETR [6], while the classification relies on CLIP.

5.1. Results

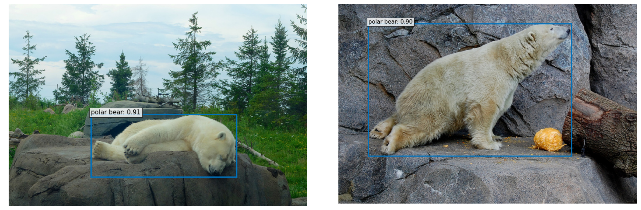
We report results on TRICD in Tab. 3 using the mean average precision (mAP) metric calculated as discussed in §3. On the VQA formulation of the task we report results in Table 4 using accuracy and macro-F1 score. We also report performance on each split, as they have different data properties and distributions (as seen in Table 1).

5.2. Discussion of results on TRICD

COCO split We break down performance on the COCO split in Table 3 in terms of the surprising object (COCO objects) and surprising relations (COCO relations) splits. We expect models that are trained on detection data to perform well on the COCO objects split, as these datasets also include negatives. The COCO relations split probes for models' understanding of relations, which is hard for models that are trained only on detection data. As per our hypothesis, MDETR, trained solely on grounding data, performs the worst on COCO objects, while DETIC, which is trained on web-scale detection data, performs the best. On the COCO relations split, we see FIBER performing the best while detection only model OWL-ViT performs the worst.

Winoground split On average the number of words in the caption per image in Winoground is 8.8 compared to 1.4 and 5.2 in COCO objects and relations, respectively. On this split, it is expected that grounding models would have an advantage and we see that indeed FIBER and GLIP-L have the best performance.

Overall, the FIBER model, which is trained in a two stage manner on image-text data and then on image-text-box data, seems to perform the best, outperforming bigger models trained on more data such as GLIP-L and object detection models like DETIC and OWL-ViT.



(a) polar bear sleeping (b) polar bear sleeping

Figure 6. Even the best performing models struggle when the verb changes between the two instances. Predictions shown here are for the FIBER model for the query phrase "polar bear sleeping": the model is insensitive to the fact that the bear in (b) is "stretching" and not "sleeping", and predicts a box with high confidence.

5.3. How discriminative is the dataset?

Given that we are proposing an evaluation benchmark, we are interested in having a measure of how well the dataset can tell apart the performance of two given models. We approximate this by randomly sampling a subset of 90% of the dataset, and evaluating the AP of our models on this subset. By repeating this process for 100 independent subsets, we obtain an estimate of the standard deviation of our metric, which we find to be around 0.5 AP. This can be considered the minimal performance gap between models that allows to conclude with confidence that a given model is better than the other. Note that in Table 3, the AP gap between any pair of models is higher than 0.5.

6. Dataset difficulty analysis

In this section we explore in more depth what makes our dataset hard for SoTA models. The CPD task can be decomposed in two sub-tasks: first a classification task to assess whether the target phrase is visible in the image, then a localization task to ground it. If a model or combination of models is able to perfectly solve both tasks, then it will perfectly solve CPD. We evaluate SOTA models on both subtasks and compare to performance on existing datasets. We show that both sub-tasks are *harder* than previously available tasks of the same nature.

6.1. Classification subtask (TRICD-VQA)

To evaluate the classification subtask, we pose it as a binary VQA task, where we ensured that all questions are well-formed.[9]

Models We use SoTA models of various sizes and scale of training data: FIBER [16], OFA [66] and Flamingo [2].

FIBER We use the coarse-grained model pretrained on 4M image-text pairs with image-text matching/contrastive and masked language modeling losses.

OFA trains a sequence-to-sequence model that is trained on image-text, grounded image-text, object detection data as

[9]For details on our question generation process see Appendix C.2.

Model	TRICD-VQA Wino	TRICD-VQA COCO objects	TRICD-VQA COCO relations	TRICD-VQA All	GQA "Exists" Testdev
Models fine-tuned on VQA					
OFA	54.3	71.7	67.7	62.0	**77.2**
FIBER	**58.5**	**75.4**	**74.7**	**66.7**	74.8
Flamingo3B	51.7	75.3	74.2	63.3	-
Model only pre-trained on general image-text data					
Flamingo80B	48.2	56.4	52.3	52.1	-

Table 4. F1 scores of SOTA models on TRICD-VQA compared to a balanced sample of "verify object" GQA questions.

well as language only data. It reformulates grounding as a sequence generation task, using ideas from Pix2Seq [10].

Flamingo uses frozen pre-trained vision and language models, and only trains adapter layers to handle sequences of arbitrarily interleaved visual and textual data. It is trained with a sequence modelling objective on web-scale data [36] and displays impressive zero shot and few shot capabilities.

Comparison dataset We compare performance on our dataset to a subset of the GQA dataset [25], one of the most challenging question answering datasets where models still lag behind human performance. None of the VQA models we report on have been trained on it, which makes it a fair zero-shot transfer performance. We filter the subset of questions in GQA that are simple yes/no questions asking about the existence of an object in the scene. There are 23185 such questions in the GQA testdev set from which we randomly sample a balanced set of 5000 total question. Note that, by design, TRICD-VQA is balanced.

Results On TRICD-VQA, FIBER achieves the best performance on all TRICD splits while all models tend to struggle most on the Winoground split. This is expected, since Winoground poses a challenge for models unable to identify when queried objects are present in the image, but not in the correct context specified by the relation. For all VQA fine-tuned models, around 60-70% of false positives occur on the Winoground split with the remaining false positives being roughly 10% more likely to come from COCO relations than COCO objects. Across the models we tested, around $50 - 60\%$ of false negatives can also be attributed to Winoground while the remaining false negatives are at least twice as likely to come from the COCO objects splits versus COCO relations. This again confirms our hypothesis that models currently under-predict the presence of surprising or out of context objects.

Compared to GQA, there is a significant gap in performance for the model we evaluated, from 8% for FIBER to 15% for OFA. This indicates that the classification subtask of our dataset is harder than previously benchmarks. Winoground is by far the most difficult split, and model performances is close to chance.

Model	TRICD-Grounding Wino	TRICD-Grounding COCO objects	TRICD-Grounding COCO relations	TRICD-Grounding All	Flickr30k Test
Grounding models					
MDETR	75.8	45.0	80.0	72.0	84.3
GLIP-T	70.6	62.7	82.2	71.7	85.7
GLIP-L	**76.2**	71.7	**86.0**	**77.5**	87.1
FIBER	74.8	68.5	85.6	76.0	**87.4**
Open vocabulary detection models					
OWL-VIT	62.3	**72.0**	78.2	66.9	-
DETIC	51.9	70.6	67.7	57.9	-

Table 5. Comparison of the grounding performance of SOTA models on TRICD and Flickr30k Entities. On both datasets, we report Recall@1 under the ANY-BOX-PROTOCOL (with IoU ≥ 0.5)

6.2. Localization subtask (Grounding)

To evaluate the localization subtask, we frame it as standard phrase grounding, which means that we *exclude any negatives* from our dataset. The models evaluated are the same as in Sec. 5.1 and we compare performance on Flickr30k Entities [54].

Metrics Following [54],we evaluate Recall@1, by using the highest confidence box for each phrase. Following the ANY-BOX-PROTOCOL [27], a box is considered correct if it has an IoU higher than 0.5 with *any* ground truth box.

Results Overall, all models evaluated have a 10% lower performance on TRICD compared to Flickr30k, indicating our grounding subtask is harder than in previous datasets. On the Winoground split, MDETR suprisingly outperforms GlipT and FIBER, despite being smaller and trained on a much smaller corpus. This split is the most challenging on the linguistic aspect, and our dataset shows that this aspect of fine-grained visio-linguistic understanding was previously a blind-spot in existing grounding datasets. On the COCO objects split, the models trained without object detection data (MDETR) are as expected struggling the most. However, even strong open vocabulary detection models such as GLIP-L and OWL-VIT obtain relatively poor performance on this set, which turns out to be the hardest for the grounding models. Finally, COCO Relations is the easiest split for this evaluation since the grounding task is comparatively easier consisting of (subject, object) pairs that involve common objects that tend to be unique in the image.

7. Conclusion

We presented TRICD, a new dataset to evaluate Contextual Phrase Detection. We believe this task is the next natural step in the quest to evaluate ever-more flexible and general detection systems. We demonstrate that the task and each of its sub-tasks (localization and classification) are challenging for current SOTA models, and we hope that this benchmark will pave the way for building stronger models with better fine-grained spatial and relational reasoning capabilities.

References

[1] Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proc. of ACL*, pages 6555–6565, Online, 2020. Association for Computational Linguistics.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint*, 2022.

[3] Maxim Tkachenko andMikhail Malyuk andAndrey Holmanyuk andNikolai Liubimov. Label Studio: Data labeling software, 2020-2022. Open source software available from https://github.com/heartexlabs/label-studio.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society, 2015.

[5] Andy Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In Marina Meila and Tong Zhang, editors, *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 1059–1071. PMLR, 2021.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[8] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1017–1025. IEEE Computer Society, 2015.

[9] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389, 2018.

[10] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society, 2016.

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of ICLR*. OpenReview.net, 2021.

[16] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *NeurIPS*, 2022.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[18] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Hartwig Adam, Matthew R. Scott, and Serge Belongie. The imaterialist fashion attribute dataset. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3113–3116, 2019.

[19] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE, 2019.

[20] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *ArXiv preprint*, abs/1505.04474, 2015.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[22] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644, Online, 2021. Association for Computational Linguistics.

[23] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *ArXiv preprint*, abs/2203.15556, 2022.

[24] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. *arXiv preprint*, 2020.

[25] Drew A. Hudson and Christopher D. Manning. Gqa: a new dataset for compositional question answering over real-world images. *CVPR*, 2019.

[26] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society, 2017.

[27] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021.

[28] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proc. of EMNLP*, pages 787–798, Doha, Qatar, 2014. Association for Computational Linguistics.

[29] Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9404–9413. Computer Vision Foundation / IEEE, 2019.

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

[33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[34] Dieu-Thu Le, Jasper Uijlings, and Raffaella Bernardi. TUHOI: Trento universal human object interaction dataset. In *Proceedings of the Third Workshop on Vision and Language*, pages 17–24, Dublin, Ireland, 2014. Dublin City University and the Association for Computational Linguistics.

[35] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*, pages 7871–7880, Online, 2020. Association for Computational Linguistics.

[36] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

[37] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022.

[38] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *ECCV*, 2016.

[39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint*, 2019.

[41] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[42] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1096–1104. IEEE Computer Society, 2016.

[43] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society, 2015.

[44] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[45] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics.

[46] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

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

[48] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh

Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *ArXiv preprint*, abs/2205.06230, 2022.

[49] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011.

[50] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision*, pages 85–100. Springer, 2016.

[51] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13018–13028, 2021.

[52] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Improving closed and open-vocabulary attribute prediction using transformers. In *European Conference on Computer Vision*, pages 201–219. Springer, 2022.

[53] Bryan Allen Plummer, Kevin Shih, Yichen Li, Ke Xu, Svetlana Lazebnik, Stan Sclaroff, and Kate Saenko. Revisiting image-language networks for open-ended phrase detection. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[54] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015.

[55] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *ECCV*, 2020.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[57] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proc. of EMNLP*, pages 4035–4045, Brussels, Belgium, 2018. Association for Computational Linguistics.

[58] Matteo Ruggero Ronchi and Pietro Perona. Describing common human visual actions in images. In Xianghua Xie, Mark W. Jones, and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015*, pages 52.1–52.12. BMVA Press, 2015.

[59] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8429–8438. IEEE, 2019.

[60] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. of ACL*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics.

[61] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[62] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proc. of ACL*, pages 217–223, Vancouver, Canada, 2017. Association for Computational Linguistics.

[63] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proc. of ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

[64] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[65] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society, 2015.

[66] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint*, 2022.

[67] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022.

[68] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[69] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *ArXiv preprint*, abs/1901.06706, 2019.

[70] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. *ArXiv preprint*, abs/2205.11169, 2022.

[71] Mark Yatskar, Luke S. Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image

understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5534–5542. IEEE Computer Society, 2016.

[72] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[73] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *ECCV*, 2016.

[74] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE, 2019.

[75] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*, 2022.

[76] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vl understanding. 2022.

[77] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. Discriminative bimodal networks for visual localization and detection with natural language queries. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1090–1099. IEEE Computer Society, 2017.

[78] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.

[79] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian D. Reid, and Anton van den Hengel. HCVRD: A benchmark for large-scale human-centered visual relationship detection. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proc. of AAAI*, pages 7631–7638. AAAI Press, 2018.

Model	TRICD Wino	TRICD COCO objects	TRICD COCO relations	TRICD All
Grounding models				
MDETR	**44.3**	30.7	50.1	43.4
GlipT	39.4	45.2	46.9	41.8
GlipL	42.4	58.5	52.3	**46.8**
FIBER	42.4	55.8	**54.1**	**46.8**
Open vocabulary detection models				
OWL-VIT	35.4	53.7	48.5	40.8
DETIC	29.2	**63.2**	40.5	36.6

Table 6. Group-Recall@1 score on subsets of TRICD

A. Alternate metric

To better analyze the performance of the models, we also report an alternative metric that we term **Group-Recall@1**. To compute it, we gather all predictions for a *positive* datapoint (I_0, C_0) as well as the predictions made for the related *negative* datapoint (I_1, C_0) where the caption is the same but the image different. Then, for each phrase of the caption, we sort all the predictions for that phrase (both those made for the positive and negative datapoints) by decreasing confidence. We consider the model successful for this phrase if the highest-confidence prediction was made for the positive datapoint, and has an IoU higher than 0.5 with any of the ground-truth boxes for that phrase.

Overall, this metric is quite similar to the Phrase Grounding metric that we report in Tab. 5. The only difference is that we consider the predictions on both the negative and positive example associated with each phrase. It tests the ability of the model to correctly rank predictions depending on whether they are positive. As such, it can be seen as a retrieval metric. Note that it doesn't evaluate any prediction beyond the top-scoring one, hence doesn't assess whether all the objects corresponding to a given phrase are detected.

The results are presented in Tab. 6. We first note that there is a 30 points gap between the Group-Recall@1 and the Phrase-grounding Recall@1. This indicates that the model gets confused pretty often by the distractor image, scoring detections higher there than in the positive one. Some significant quantitative differences between the AP results (Tab. 3) and the Group-Recall@1 can be observed. The most striking one is the performance of MDETR on the Winoground split: according to the AP metric, it performs significantly worse than all the other grounding models, while according to the Group-Recall@1 metric it performs the best. This indicates that MDETR has a better *intra-phrase* calibration (it tends to rank positives higher than negatives for a particular phrase), but overall worse *inter-phrase* calibration (at the dataset level, positives and negatives do not get ranked correctly, leading to poor AP).

B. Validation set

Stats	Coco obj	Coco relation	Overall
Unique images	40	60	99
Unique phrases	43	73	114
Unique words	64	131	188
image/caption pairs	84	120	204
average phrase/img	1.0	2.0	1.6
average box/phrase	1.9	1.9	1.9
Total number of boxes	83	232	315
Average words/caption	1.6	5.0	3.6

Table 7. Statistics of the TRICD validation set.

Model	TRICD-val COCO objects	TRICD-val COCO relations	TRICD-val All
Grounding models			
MDETR	6.9	17.8	14.1
GLIP-T	22.8	22.6	21.4
GLIP-L	**30.2**	26.3	**26.0**
FIBER	19.6	**28.2**	25.8
Open vocabulary detection models			
OWL-VIT	9.6	12.0	11.2
DETIC	19.9	21.9	20.4

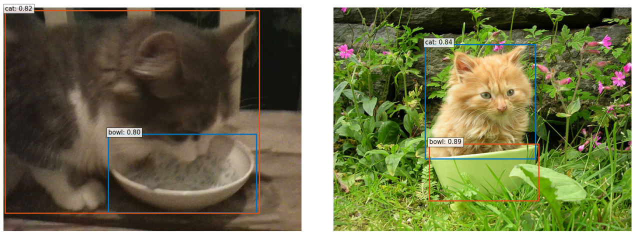
Table 8. Average Precision (AP) score on the validation subsets of TRICD

To ease experimentation on our dataset, we provide an additional validation set, only for the coco split of our data. The annotation procedure is exactly the same as the coco split of our test set, except the images come from the Val2017 subset of coco. As a result, there may be some overlap with some training sets of other datasets based on COCO (eg LVIS). We report statistics of our validation dataset in Tab. 7. Results on the CPD task are reported in Tab. 8.

C. Details on the evaluation

C.1. Inference parameters

For evaluation on CPD, we ensure that all the models predict at least 100 bounding boxes per image for calculation of the AP metric. For MDETR, DETIC, and OWL-ViT, default configurations are sufficient. For GLIP-L, GLIP-T, and FIBER, the all post-processing thresholds must be set to 0 and config parameter $\mathrm{MODEL.ATSS.PRE\_NMS\_TOP\_N} = 3000$. For GLIP-L and GLIP-T the following config parameters must



(a) cat next to a bowl (b) cat next to a bowl

Figure 7. Additional example of a miss-prediction by FIBER. Here, the model is insensitive to the attribute ”next to” and produces high confidence detection in the second image, even though the correct attribute in this case is ”inside”.

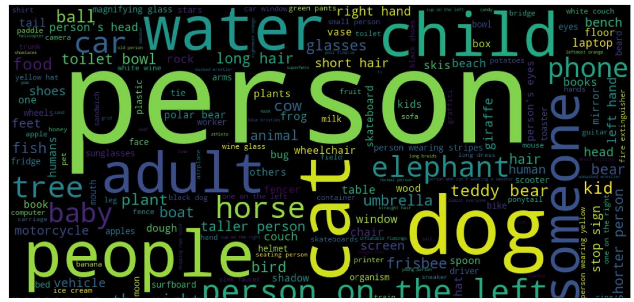


Figure 8. Word cloud for phrases

also be set: MODEL.ATSS.INFERENCE_TH = 0, and MODEL.ATSS.NMS_TH = 0.6.

C.2. Converting captions to VQA format

Captions are converted into questions using the NLTK [45] and Inflect packages for part of speech (POS) tagging, followerd by manual verification. For the Winoground split, a mixture of common pattern matching (i.e. a sentence beginning with “there is” can usually be converted into a question by simply switching the word order to ”is there”) and POS tagging was used. However, given the complexity of some Winoground phrases, it was necessary to manually generate custom questions for 184 out of total phrases. For the COCO split, since many phrases are a single word or short phrase, it is straightforward to systematically convert these into questions. A few question words are applied based on the POS of the first word in the sentence. For instance, if the first word is a singular noun, the question is formed as “Is there a” + phrase+“?”. If the first word is an article (“a”), the question would be generated as “Is there” + phrase + “?”. Additional manual verification for grammatical correctness was applied for both sets.

D. Dataset analysis

In Fig. 8, we give a glimpse of the content of the dataset by computing a word clould of the individual phrases.

E. Annotation process

E.1. AWS annotation details

The Winoground images are relatively easy to understand (stock images from Getty Images API). We set the price per HIT to \$0.048 as suggested by SageMaker for a job that takes 11-13 seconds. We also run a separate job for images that are difficult to understand or contain many objects to be annotated per phrase, where the price per HIT is increased to \$1.20 as they are expected to take between 3 and 3.5 minutes. Given that COCO images come from a different data distribution, having complex scenes, many of the examples contain phrases that are difficult to find in the image and/or obscure long-tailed concepts. We set the price of the HIT to \$0.24 for an estimated time of 23-25 seconds per image.

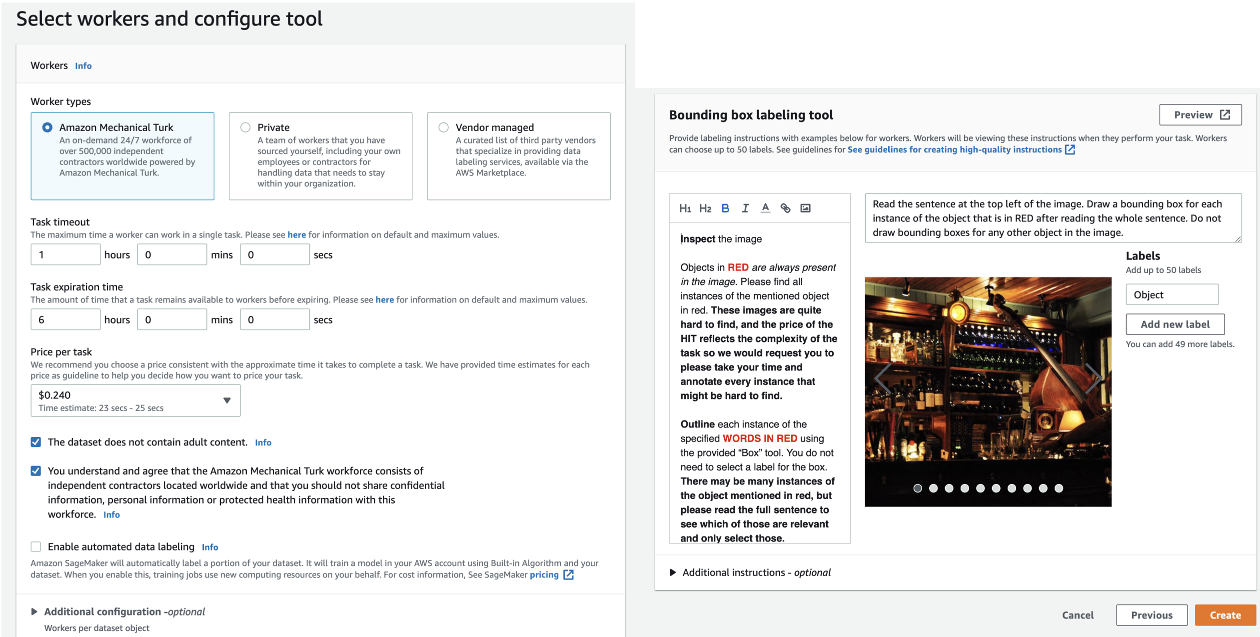
E.2. Annotation Tool

Screenshots for the annotation interface along with the instructions and settings are shown in Figs. 9 to 12.

F. Details on related datasets

Phrase Grounding The Flickr30k Entities dataset [54] consists of 30,000 images annotated with 5 captions each, where for each noun phrase in the caption, an associated set of bounding boxes is provided. In Phrase Grounding, success is defined in terms of whether the model predicts a box with an intersection over union (IoU) of at least 0.5 with the target box for each phrase in the dataset. The IoU threshold of 0.5 is chosen in part because of the inherent noise in the annotations that prevents much more stringent metrics. The metric that is commonly used to evaluate performance on this task is the Recall @k metric, with k = 1, 5 and 10 being the slack in the number of boxes that the model can predict before predicting the correct box (when ranked in terms of confidence). An important point to note, is that during the task of Phrase Grounding, it is assumed that the phrases being queried do necessarily exist in the image. Current state of the art models such as MDETR [27], GLIPv1 [37], GLIPv2 [76], FIBER [16] and PEVL [70] have come close to just 10% error rates on this dataset. [10] While this could imply that these models have extremely good grounding abilities, in reality we find that when queried with negative phrases, the models perform terribly, leaving much to be explored in the direction of models that possess true visual understanding abilities.

[10] which has been reported to be close to the upper bound according to analysis on dataset noise carried out in [27]



(a) Worker wages and task timeout settings

(b) Set up for the annotation interface

Figure 9. Illustration of the settings for the bounding box annotation tool (Sage Maker)

Referring expression comprehension (REC) The REC task involves returning a bounding box for each referring expression that uniquely identifies an object, given an image. When predicting a bounding box, the model has to consider the relative spatial information of other objects in the image of the same type as well as make visual comparisons to similar objects, to disambiguate between them. This probes the model's attribute and spatial understanding abilities. RefCOCO, RefCOCO+ [28] and RefCOCOg [73] are large scale datasets collected on natural images from the COCO dataset [39] with on the order of 100k expressions per dataset. The metric that is used to evaluate on this task is the accuracy at IoU threshold 0.5, where a true positive is defined as a bounding box that has at least IoU 0.5 with the target ground truth box for each referring expression. Current state of the art VL systems have close to 90% accuracy on these datasets. The expressions used to describe the objects are often limited in vocabulary and very short in length. The Referring Expression Generation task is the converse task of predicting a natural language desription given an image and a bounding box, and common metrics of evaluation are BLEU, METEOR and ROUGE. Ref-Adv [1] is a more recent adversarial split of the RefCOCOg dataset which probes for the model's sensitivity to word order in the referring expression.

Winoground Closely related to the topic of models being insensitive to word order, is the Winoground dataset [64] consisting of 800 unique captions and images. Here the goal is to match the correct pairs given two images and two captions on this dataset having 800 correct and 800 incorrect pairings. The difficulty of this task lies in the fact that the two captions use the same set of words, but differ in word order. In a subset of the dataset, the two images are also taken from the same scene which further challenges models trying to discriminate the correct pairs. The metrics used by [64] to measure such visio-linguistic compositional reasoning are image score, which measures whether a model can select the correct image, given a caption and text score which measures the converse. They also use a group score which takes into account both the of the previous scores. Most SOTA models currently perform barely better than chance on this dataset. While proposed as a fine-grained visual understanding task, the matching of images and text provides limited signal in uncovering the models ability to understand the complex compositional reasoning required to solve this task, which inherently requires knowledge of objects and their relations. In Sec §4 we describe our proposal to more deeply evaluate the models for compositional reasoning though our dataset.

Attribute Prediction The VAW dataset [51] consists of 72,274 images from the Visual Genome dataset annotated with 620 unique attributes for over 260k object instances, that represent a long tail of object-attribute pairs superseding previous attempts in terms of size and coverage. Differently from the phrase detection dataset [53], VAW is a federated dataset that provides certificates for negatives per

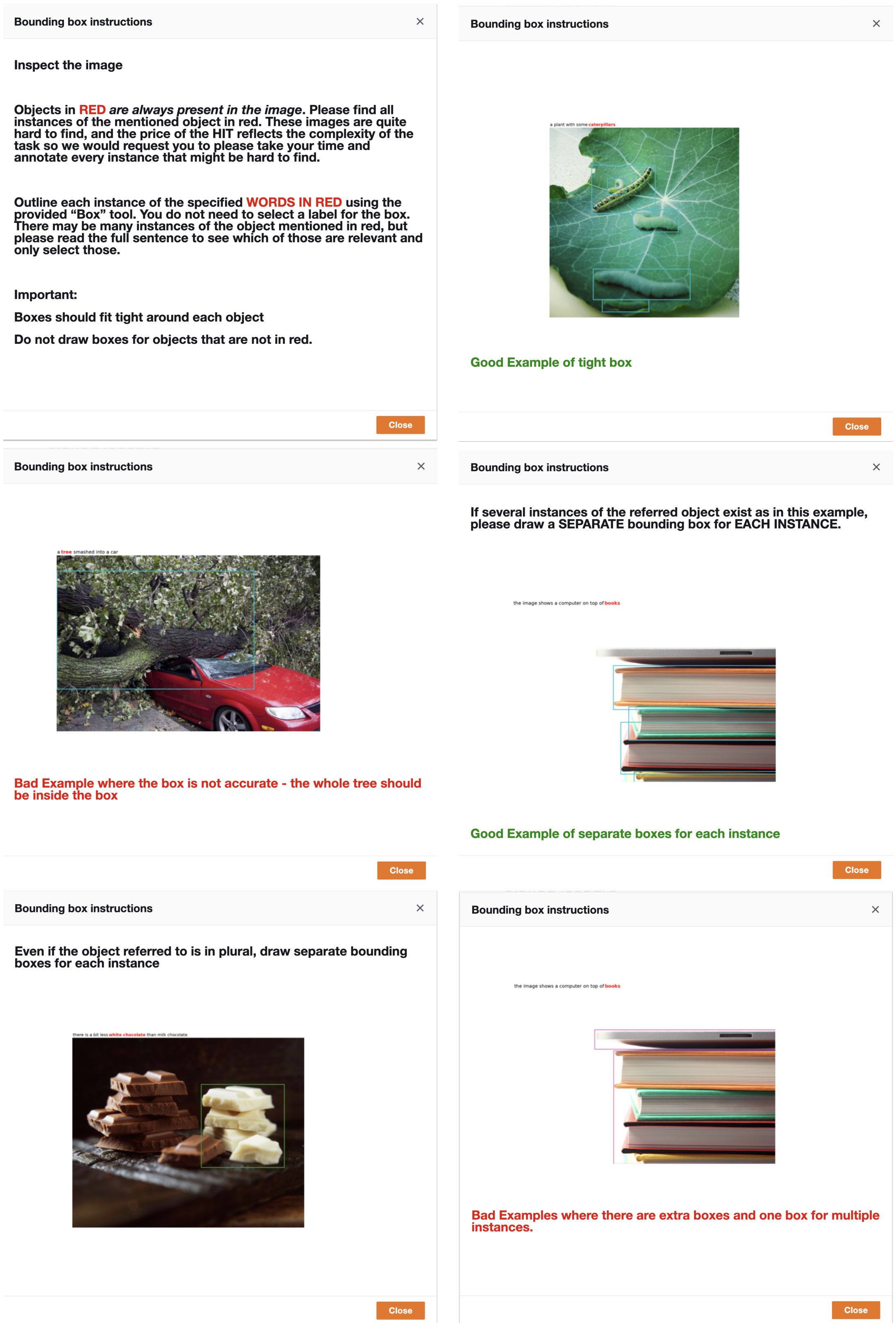


Figure 10. Illustration of the instructions provided to workers for bounding box annotation.

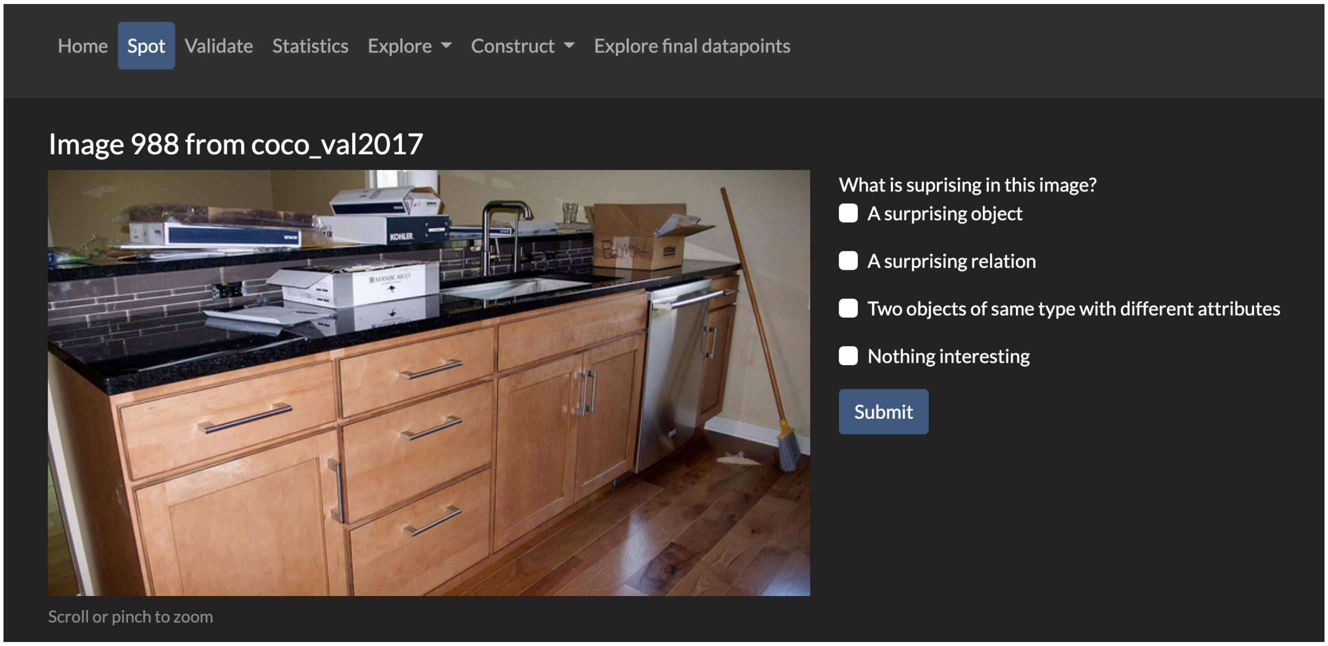


Figure 11. Interface of the spotting tool, where annotators must flag images which have interesting objects or relations

object. This allows for accurate evaluation of the models ability to predict the presence or absence of each attribute, taking only into account the relevant positive and negative objects per attribute. More recently, the LSA dataset [52] combines images from more sources such as Flickr30k [72], COCO [39] and OpenImages [33] to create a larger visual attribute detection dataset

Relation Prediction In addition to attributes, another important capability of visual understanding models is the ability to recognize relations. HICO [8] and VRD [46] are relation prediction datasets which involve classifying the detected relationship. In HICO the task is to classify the interaction of a human with an object, and VRD requires classification of the relationship between two objects. Both of these have a limited set of verbs and objects. The SVO-Probes dataset is an evaluation benchmark having 48,000 image–text pairs designed to probe for Subject, Verb, and Object understanding in image-text models. It consists of image-text pairs covering 421 verbs that are considered to be visual and extracted from the Conceptual Captions dataset [60]. The difference between the positive and negative image is either in the subject, verb or object and the task is to correctly classify both positive and negative pairs. The performance of SoTA models on this dataset suggests that models struggle on verbs, as compared to recognizing other parts of speech. Other datasets such as V-COCO [20] and ImSitu [71] probe for verbs but not with negative certificates as in SVO-Probes.

G. Models used for evaluation

MDETR is an end-to-end object detection pipeline built on DETR [6] and conditioned on free form text. It predicts bounding boxes and which words in the input caption they correspond to. MDETR predicts a set of bounding boxes given an image and a text query, as well as a distribution for each predicted box over the tokens of the input text used to query the model. We evaluate MDETR-ENB5 which has an EfficientNet-B5 vision backbone and RoBERTa as the text encoder. It is trained on 1.3M image-text pairs from COCO [39], VG Regions [30], GQA [25] and Flickr30k [54], together referred to as GoldG. This model has not been trained on negative examples (e.g. through object detection data) and hence is expected to perform poorly on the negatives in our dataset.

GLIP casts object detection as a grounding task and incorporates both kinds of data in its training. The GLIP-L model that we evaluate is trained on data including 4 object detection datasets (Objects365 [59], OpenImages [33], Visual Genome [30] and ImageNetBoxes [32], 24M pseudo-annotated image-text pairs from the web, CC12M [7], SBU captions [49], as well as GoldG from MDETR. GLIP-L uses a Swin-Large [41] as the vision backbone and BERT as the text encoder. GLIP-T is trained on GoldG and Objects365 and uses a Swin-Tiny as the vision backbone.

FIBER extends GLIP and leverages coarse-grained image-text pre-training for subsequent fine-grained image understanding by having a fused backbone architecture that fuses the image and text modalities deeper in the model

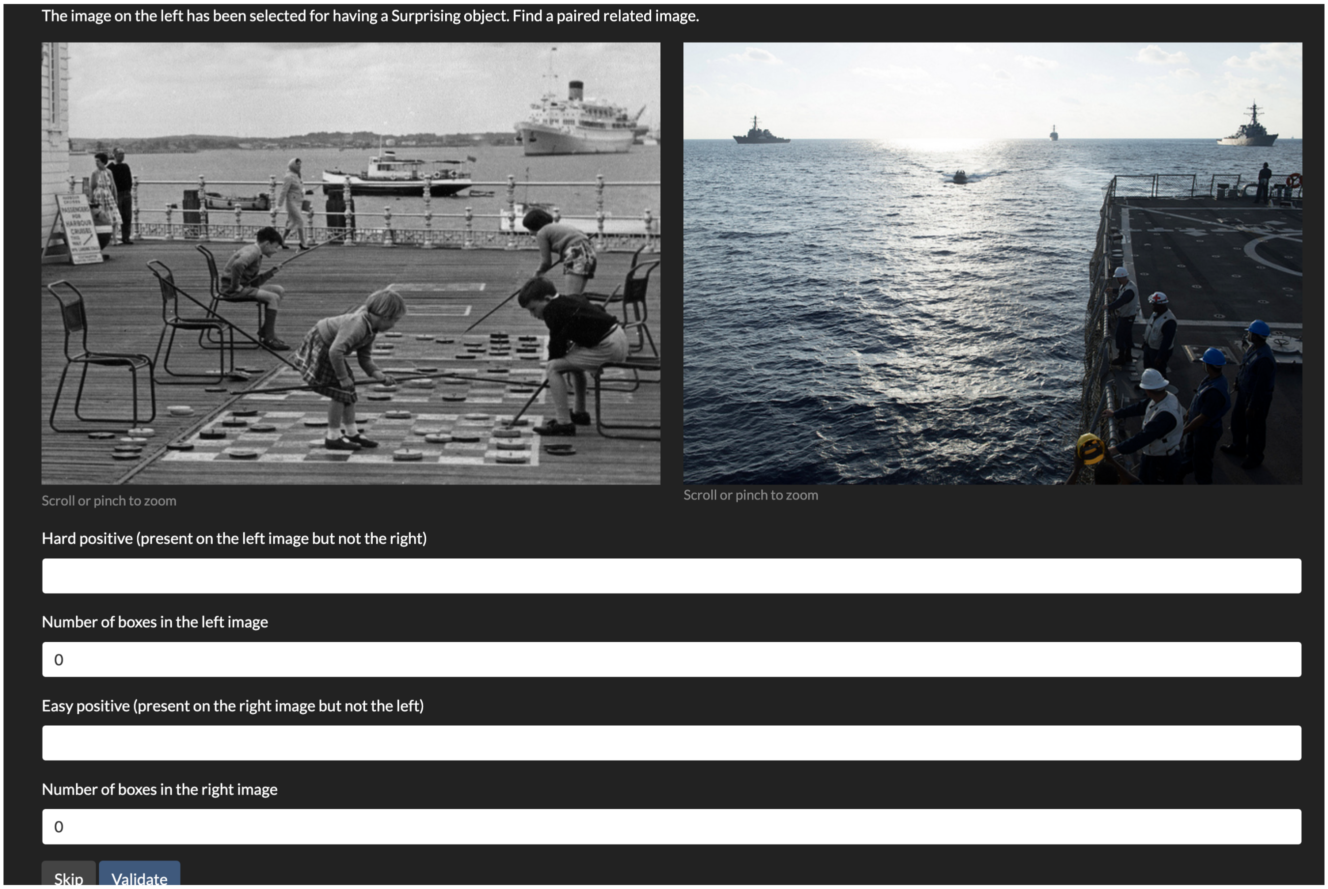


Figure 12. Interface of the construction tool. Given an image an a caption, annotators must find a related image where the given caption does not occur, and construct a positive caption for this second image that conversely does not occur in the first image

compared to MDETR or GLIP. This allows a large amount of parameters to be initially trained on image-text data, providing a good initialization for the fine-grained training of the model and reducing the requirement for box annotated data. The model we use is based on Swin-Base [41] and RoBERTA [40], and is trained on the Gold-G data from MDETR, object detection data from Objects365 [59] as well as image-text pairs from COCO, VG [30], CC3M [60] and SBU captions [49].

DETIC is an open-vocabulary detector that uses CLIP [56] embeddings to encode the class names. It leverages a mixture of box-annotated data as well as image-level annotations from ImageNet, with a weakly-supervised loss.

OWL-ViT also relies on CLIP, relying on a very large VIT [15]. During fine-tuning, it uses object detection datasets to train a localization head, using a matching loss similar to DETR [6], while the classification relies on CLIP.

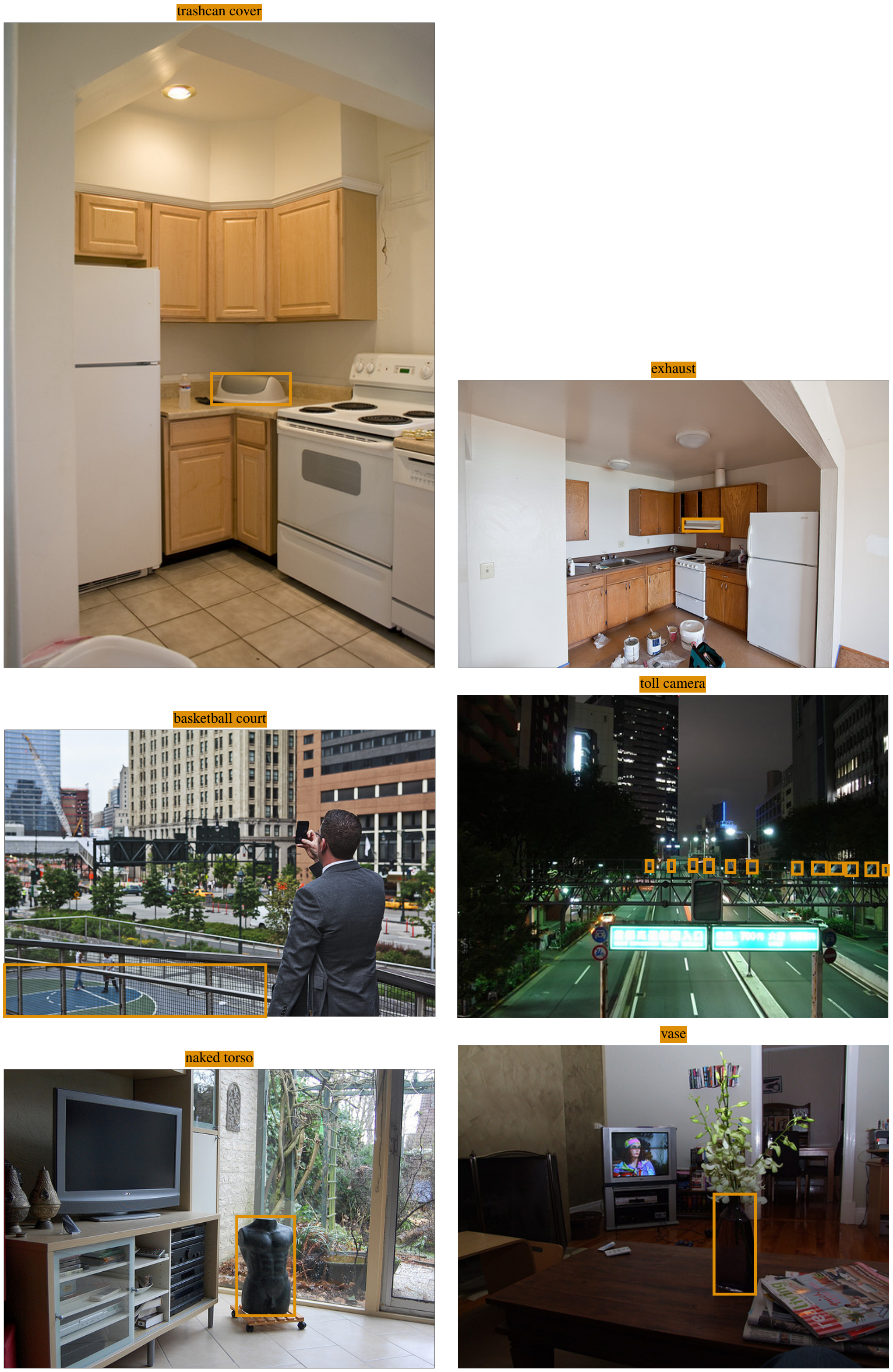


Figure 13. Random examples from the object split of the val dataset

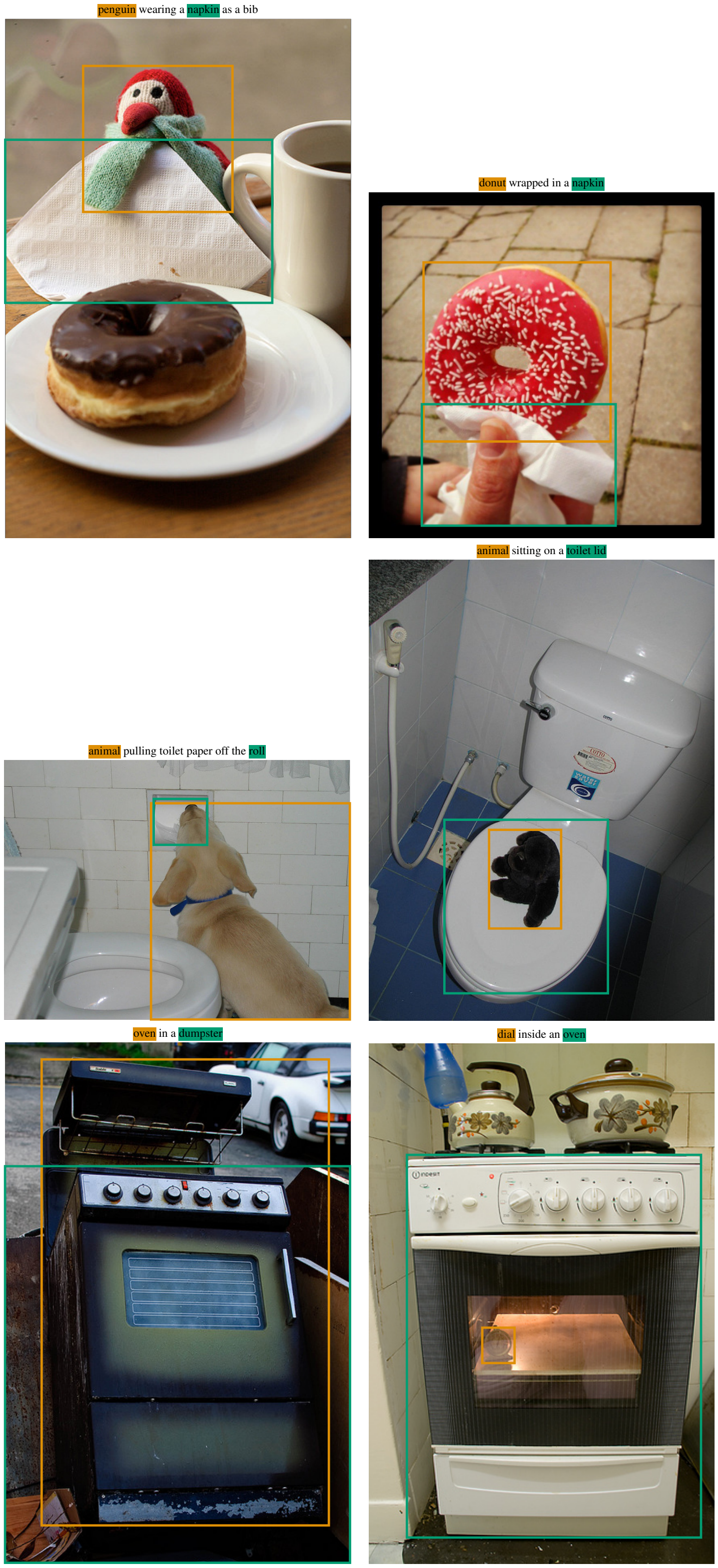


Figure 14. Random examples from the relation split of the val dataset