

Creative Commons Attribution 4.0 International (CC BY 4.0)

<https://creativecommons.org/licenses/by/4.0/>

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

To Find Waldo You Need Contextual Cues: Debiasing *Who's Waldo*

Yiran Luo Pratyay Banerjee Tejas Gokhale Yezhou Yang Chitta Baral

Arizona State University, Tempe, AZ, USA

{yluo97, pbanerj6, tgokhale, yz.yang, chitta}@asu.edu

Abstract

We present a debiased dataset for the Person-centric Visual Grounding (PCVG) task first proposed by Cui et al. (2021) in the *Who's Waldo* dataset. Given an image and a caption, PCVG requires pairing up a person's name mentioned in a caption with a bounding box that points to the person in the image. We find that the original *Who's Waldo* dataset compiled for this task contains a large number of biased samples that are solvable simply by heuristic methods; for instance, in many cases the first name in the sentence corresponds to the largest bounding box, or the sequence of names in the sentence corresponds to an exact left-to-right order in the image. Naturally, models trained on these biased data lead to over-estimation of performance on the benchmark. To enforce models being correct for the correct reasons, we design automated tools to filter and debias the original dataset by ruling out all examples of insufficient context, such as those with no verb or with a long chain of conjunct names in their captions. Our experiments show that our new sub-sampled dataset¹ contains less bias with much lowered heuristic performances and widened gaps between heuristic and supervised methods. We also demonstrate the same benchmark model trained on our debiased training set outperforms that trained on the original biased (and larger) training set on our debiased test set. We argue our debiased dataset offers the PCVG task a more practical baseline for reliable benchmarking and future improvements.

1 Introduction

A newly released task called Person-centric Visual Grounding (Cui et al., 2021) poses an interesting angle into contextual reasoning in vision-language. The task is motivated by humans' reasoning ability.

¹Available at: <https://github.com/fpsluozi/tofindwaldo>



Figure 1: We find many biased data from the original *Who's Waldo* dataset contain insufficient contextual cues and cannot be used to map names to persons in an image. **Left:** An unsolvable example with no actions nor descriptions w.r.t the detected persons. Given no background knowledge about the individuals, one can only guess the masked [NAME]'s based on heuristic biases such as the locations of the bounding boxes. **Right:** A qualifying example with clearly worded interactions (e.g. detectable verbs such as 'watches' & 'signs') about each masked name - the very type of data we incorporate into our debiased dataset.

Humans viewing an image with a caption as shown in Figure 1 can reason (and if needed, speculate) which name refers to which person in the image. This reasoning task involves multiple abilities, such as perceiving characteristics and behaviors of people, understanding their actions in context, speculating about their intentions and effects human of actions (Fang et al., 2020), and connecting visually perceived characteristics with grounded descriptions in natural language (Kazemzadeh et al., 2014; Yu et al., 2016; Zellers et al., 2019). In many cases, this task can be performed without knowing the names of the people; for instance in the example on the right, one person is signing and the other is not, as such it is possible to predict which person refers to President and Secretary of State respectively. However, in cases such as the example on the left, if all persons are performing the same action (run-

ning on a track), then it is hard to match names with these runners without any additional information. Progress in the PCVG task can thus help better capture what exact contextual cues are needed to learn about a person’s characteristics in a scenario, and can aid improvements in visual understanding about human interactions and behaviors.

To support this task, Cui et al. (2021) offer a large-scale dataset called *Who’s Waldo* which consists of 272K annotated real life images. Ideally, the dataset should consist of input-output pairs (such as the example on the right in Figure 1) which are ‘solvable’ as opposed to the one on the left which is ambiguous. However, as we explore the original *Who’s Waldo* dataset, we encounter a great portion of cases that resemble the left example in Figure 1, unsolvable data with insufficient contextual cues. Given such context, if we do not recognize who exactly is in the picture, even we human beings cannot tell which name is who. We can then only make predictions with biased assumptions, such as the first named person would always be on the leftmost, or the main subject would always make up the largest area. Such biases in the original dataset may explain why the heuristic methods perform very strongly, outperforming random guessing by a big 27% increase in test accuracy and trailing the top benchmark only by 6%. We believe a fair dataset should not encourage approaches to adopt biases to such an extent, and thus the original baseline model overestimates its performance.

Inspired by dataset debiasing works such as VQA-CP (Agrawal et al., 2018) and GQA-OOD (Kervadec et al., 2021), we create a debiased collection of 84K annotated image-captions out of the *Who’s Waldo* dataset by filtering out all biased data with insufficient context. We evaluate the quality of our new dataset by applying the original heuristic methods as well as *Who’s Waldo*’s benchmark model. Results show that our debiased dataset greatly reduces the heuristic biases from the original dataset and provides the PCVG task a more practical baseline for future developments.

2 Related Work

Dataset Debiasing. We take many inspirations from previous studies on uncured datasets. A task dataset if not curated properly could lead to methods that cheat their ways through without learning generalized information. For example, VQAv2 (Goyal et al., 2017) addresses the imbalance be-

tween language and images in VQAv1 (Antol et al., 2015) which results in visual information being ignored and inflated model performance. VQA-CP (Agrawal et al., 2018) and GQA-OOD (Kervadec et al., 2021) were designed to test model performance if spurious correlations exist in the training dataset. Cadene et al. (2019); Chen et al. (2020a); Gokhale et al. (2020) are bias-aware techniques that mitigate dataset bias with modeling and data augmentation. Ye and Kovashka (2021) introduce exploits by matching repeated texts in questions and answers to achieve high scores in Visual Commonsense Reasoning (Zellers et al., 2019).

We also learn from various techniques to amend priors, biases, or shortcuts in datasets. REPAIR (Li and Vasconcelos, 2019) uses resampling to fix representation biases in image datasets. Dasgupta et al. (2018) incorporate compositional information into sentence embeddings for Natural Language Inference. DQI (Mishra et al., 2020) offers quantitative metrics to assess biases in automated dataset creation in Natural Language Processing. Le Bras et al. (2020) introduce adversarial measures to mitigate biases in various Natural Language Processing and Computer Vision tasks.

Visual Grounding. The PCVG task adapts previous supervised Visual Grounding models as its original baselines. The Visual Grounding task is defined as locating specific objects in an image from a textual description. First established by Karpathy et al. (2014), following researches have evolved into extracting attention information such as works by Deng et al. (2018) and Endo et al. (2017). A huge variation of datasets for Visual Grounding have also been created, including Flickr30k (Plummer et al., 2015), Visual Genome (Krishna et al., 2017), and RefCOCO (Yu et al., 2016).

Referring Expression Comprehension (REC). An active branch from Visual Grounding, the Referring Expression Comprehension task (Rohrbach et al., 2016) is no longer restricted to object categories. Instead its goal is to relate a free region in an image to a sentence description. Mattnet (Yu et al., 2018) is one prominent approach that leverages both attention features and relation extraction for the objects in the image. Qiao et al. (2020) offers a comprehensive survey on this topic.

Human Detection. A specialized category under Object Detection, detecting humans with bounding boxes in images nowadays can easily use open source toolboxes including MMDetection (Chen

Selection	Train	Val.	Test	Unused
Original	179073	6740	6741	79193
no-verb	125585	3446	3529	34366
conjunct-names	16446	2237	2227	15693
Ours	45884	2102	2049	33611

Table 1: The data in our debiased dataset are filtered and regrouped from all four splits in the original. Notice, examples such as the **Left** in Figure 1 can have both zero verb and at least three conjunct names.

et al., 2019) or Detectron (Wu et al., 2019) that are trained on large-scale real life image datasets like COCO (Lin et al., 2014). Recent works such as DarkPose (Zhang et al., 2020) also attempt to utilize human pose information to better single out human traits from complex background.

3 Method

In this section, we introduce the Person-centric Visual Grounding task, discuss the original *Who’s Waldo* dataset, and provide our analysis of shortcuts, biases, and other issues that we discovered in the dataset. We describe the process via which we curate, debias, and filter the dataset.

3.1 The Task

The Person-centric Visual Grounding task is defined as follows. The givens are an image \mathbf{I} , a set of $m \geq 1$ person detections \mathbf{B} (in form of bounding boxes), and a corresponding image caption \mathbf{T} where its tokens contain references to $n \geq 1$ persons. For each referred person, we look for the best matching detection from the givens. We also assume no two persons can be matched with the same detection.

3.2 The *Who’s Waldo* Dataset

The dataset consists of 272K real-life captioned images sourced from the free Wikimedia Commons repository. Each image pictures individuals under the ‘People by name’ category on Wikimedia Commons, while its caption describes the scene and explicitly mentions the featured people in real names. Key dataset creation procedures, text pre-processing, identifying person entities in captions, detecting bounding boxes of people in images, and generating ground truths linking bounding boxes and names, are all done with existing automated tools such as FLAIR (Akbik et al., 2019) and MMDetection (Chen et al., 2019). To prevent misuse, in the publicly released version, all the

real names in the captions are replaced with the [NAME] token, but references between bounding boxes and token indices are given in individual annotation files. This is equivalent to masking each name with indexed placeholders such as PERSON1, PERSON2, etc. Amongst the entirety of 272K annotated samples, 179K samples are used for training, 6.7K for validation, and 6.7K for testing. Each test sample is supposed to either *mention at least two persons* or *choose from at least two bounding boxes*. The original test set is further validated manually on Amazon Mechanical Turk.

3.3 Biases in *Who’s Waldo*

The premise of the Person-centric Visual Grounding task is to use ONLY the caption text and the image as the cues to find out the correct bounding box from the image per mentioned name. However, we observe a large portion of the original *Who’s Waldo* dataset does not provide sufficient contexts and can only be solved by heuristic methods. We discuss two major types of biases that we discover in the following sections.

The first type `no-verb` is that the caption text contains zero detectable verbs. Since linguistically a verb is the crucial part of an action that assigns participants with semantic roles, we technically have no way to tell who performs or who receives an action without verbs. For example in Figure 2(a), we are unable to tell who is who from the image and the no-verb caption alone, unless we recognize Vladimir Putin or the Georgian President with external knowledge.

The second type `conjunct-names` is that the caption contains a long chain of conjunct referred names. Shown in Figure 2(b), all the referred names share the verb *perform*, joined together only with conjunct words such as *and* or *along with*. With no indication of the order amongst these persons, we can only resort to a naive positional order such as left-to-right. But since we may also have extra bounding boxes as choices, such naive assumption is indeed unreliable. Figure 2(b) is such an example that the first mentioned name is not always the one in the left-most bounding box.

3.4 Data Curation for De-biasing

In order to resolve the aforementioned limitations of the original dataset, we utilize two pipelines in SpaCy ver 3.0 (Honnibal et al., 2020) to filter out the biased data. We apply the POS-Tagging pipeline to find out if sentences in an image cap-

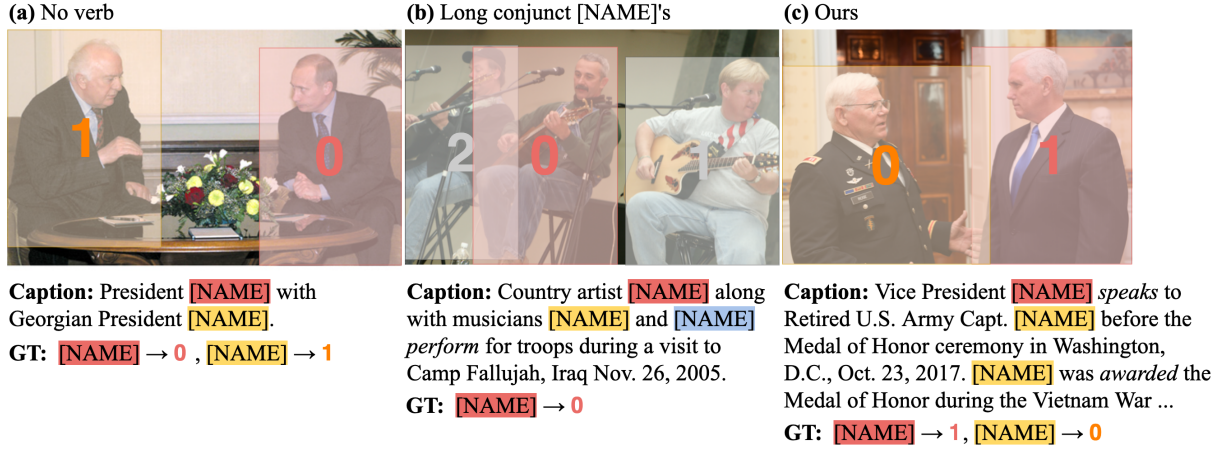


Figure 2: (a) and (b) represent the two major types of insufficient and biased data that we filter out. (c) represents the ones we choose for our debiased dataset. We label all detected verbs in *italic*. We apply color coding to indicate different person entities in a caption. We also use gray bounding boxes to refer to those 'incorrect options' not included in ground truth, such that in the ground truth of (b), the only pair we need to associate is [NAME] with Bounding Box 0, while the two other bounding boxes serve as mere distractions.

tion contain verbs in any form of conjugation. In parallel, we use the Dependency Parsing pipeline to examine if any [NAME] token conjuncts with more than one [NAME]'s from different referred persons. We jointly filter out any example that either (a) contains zero verbs, or (b) has at least three conjunct referred person names in a sentence. For both pipelines, we replace the [NAME] tokens that refer to the same person in a caption with a random popular first name, so that the natural language-based SpaCy pipelines can yield more accurate results. Both pipelines use the state-of-the-art `en-web-core-trf` model which is built on RoBERTa (Liu et al., 2019).

Ultimately, our filtering procedure produces 84K qualifying image-caption pairs. Table 1 shows the distribution of samples sourced from each split of the original through our two debiasing pipelines. We utilize data from the unused yet legitimately annotated 79K samples of the original dataset. We reorganize and split all the qualifying 84K samples into 74K for training, 5K for validation, and 5K for test. Our new test set does not overlap with the original training set. Similarly to the design of the original, we enforce that all samples in our new test set involves no trivial case that contains exactly one referred name and exactly one bounding box. We also make sure that any test set sample always has at least one name-to-bounding-box pair as ground truth.

4 Experiments and Baselines

Setup. We evaluate the quality of our debiased dataset with the same heuristic and Transformer-based methods from the original paper. We also train the benchmark model on both the original and our new training set. We report the accuracies obtained from our new test set as the new baselines.

Heuristics. We inherit the original heuristic measures to study the potential biases of our debiased dataset versus those of the original dataset. Alongside Random guessing, we assign the names in the caption to the bounding boxes sorted by: (a) decreasing area size (Big → Small), (b) left-to-right upper-left coordinates (L → R (All)), and (c) left-to-right upper-left coordinates of the largest d bounding boxes, d being the larger between the number of bounding boxes and the number of names in a test case (L → R (Largest)).

Transformer-based Models. We adapt the original benchmark *Who's Waldo* model to our debiased dataset and see how well it can perform under the updated contexts. The benchmark model is a multi-layer multi-modal Transformer (Vaswani et al., 2017). Based on UNITER (Chen et al., 2020b), it learns to maximize the similarities between the corresponding person names and bounding boxes while minimize the similarities between those that do not match up. We fine-tune the *Who's Waldo* model with pre-trained weights from UNITER.

Analysis of Results. Table 2 shows the test set accuracies for the original dataset and our debi-

Method	Training Set	Test Set	Test Accuracy	Δ_r	Δ_h
Random	–	Original Test	30.9	0.0	–
Big \rightarrow Small	–	Original Test	48.2	+17.3	–
L \rightarrow R (All)	–	Original Test	38.4	+7.5	–
L \rightarrow R (Largest)	–	Original Test	57.7	+26.8	0.0
Gupta et al.	COCO	Original Test	39.3	+8.4	-18.4
SL-CCRF	Flickr30K Entities	Original Test	46.4	+15.9	-11.3
MAttNet	RefCOCOg	Original Test	44.0	+13.1	-13.7
<i>Who's Waldo</i>	Original Train	Original Test	63.5	+32.6	+5.8
Random	–	Our Test	31.0	0.0	–
Big \rightarrow Small	–	Our Test	43.8	+12.8	–
L \rightarrow R (All)	–	Our Test	32.4	+1.4	–
L \rightarrow R (Largest)	–	Our Test	44.3	+13.3	0.0
<i>Who's Waldo</i>	Original Train	Our Test	50.2	+19.2	+5.9
<i>Who's Waldo</i>	Our Train	Our Test	54.0	+23.0	+9.7
<i>Who's Waldo</i>	Our Train	<i>Biased samples</i> of Original Test	48.2	–	–

Table 2: Evaluation on the test sets using the original *What's Waldo* and our debiased dataset. Δ_r denotes relative improvement over random guessing, and Δ_h denotes relative improvement over the best heuristic. The *biased samples* represents a total of 4.7K samples from the original test set that are filtered out by our debiasing procedure. The original work also compares its baseline performance with multiple pre-trained visual grounding models, such as Gupta et al. (2020) trained with COCO (Lin et al., 2014), SL-CCRF (Liu and Hockenmaier, 2019) trained with Flickr30K Entities (Plummer et al., 2015), and MAttNet (Yu et al., 2018) trained with RefCOCOg (Mao et al., 2016). All reported accuracies in this table are the strongest averaged performances per setting and fall within a fluctuation of $\pm 1\%$.

ased dataset. We find that the heuristic measures have overall lower performance on our new dataset, meaning we have successfully reduced the effects of the positional and the size-based biases from the original dataset. Most significantly, we have lowered L \rightarrow R (All) from +7.5% to +1.4%, almost equal to randomness. Even the strongest L \rightarrow R (Largest) heuristic has been lowered from +26.8% all the way down to +13.3% as well. Our dataset is thus proven less biased compared to the original.

We also show that our dataset has better practicality for the task. Measured with our new test set, the performance of the *Who's Waldo* benchmark model trained with the original training set performs 3.8% lower than that trained with our new, smaller training set. Meanwhile, the test accuracy gap between the Transformer-based method and the heuristic methods has become larger using our debiased dataset, widened from 5.8% to 9.7%. In addition, using the filtered *biased samples* from the original test set on our new trained model yields an even lower performance at 48.2%, which indicates our new baseline model now adopts fewer biases during training compared to the original. Altogether with the lowered new baseline accuracy of 54.0%, we argue that our debiased dataset improves the quality of contextual cues that su-

pervised models can learn from, and leaves more applicable room for improvements in the future.

5 Conclusion

We present a refined dataset for the PCVG task with samples that contain contextual information required for the task. We address prominent biases that we identified in the original task dataset by filtering out a large number of unsolvable cases, and report new baseline performances on the new benchmark. Our refined dataset can serve as a more reliable benchmark to enable fair comparisons for new modeling techniques and training protocols.

Acknowledgements

This research was supported by grants from DARPA SAIL-ON, DARPA KAIROS, NSF 1816039, and NSF 2132724. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

Ethical Considerations

Our curated dataset is available at <https://github.com/fpsluozi/tofindwaldo>. We will also follow the same licensing and data sharing policy as the original *Who's Waldo* dataset.

References

- Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.
- Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32.
- Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. 2019. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.
- Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020a. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. pages 104–120.
- Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snaveley, and Hadar Averbuch-Elor. 2021. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1374–1384.
- Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.
- Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. 2018. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755.
- Ko Endo, Masaki Aono, Eric Nichols, and Kotaro Funakoshi. 2017. An attention-based regression model for grounding textual phrases in images. In *IJCAI*, pages 3995–4001.
- Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Video2commonsense: Generating commonsense descriptions to enrich video captioning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 840–860.
- Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.
- Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spacy: Industrial-strength natural language processing in python](#).
- Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.
- Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. 2021. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2776–2785.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.
- Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

- Yi Li and Nuno Vasconcelos. 2019. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9572–9581.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Jiacheng Liu and Julia Hockenmaier. 2019. Phrase grounding by soft-label chain conditional random field. pages 5115–5125.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.
- Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. 2020. Dqi: Measuring data quality in nlp. *arXiv preprint arXiv:2005.00816*.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.
- Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*.
- Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. <https://github.com/facebookresearch/detectron2>.
- Keren Ye and Adriana Kovashka. 2021. A case study of the shortcut effects in visual commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3181–3189.
- Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315.
- Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.
- Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. 2020. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102.