Run Like a Girl!
Sports-Related Gender Bias in Language and Vision

Sophia Harrison
Universitat Pompeu Fabra
sophia.harrisonn@gmail.com

Eleonora Gualdoni
Universitat Pompeu Fabra
eleonora.gualdoni@upf.edu

Gemma Boleda
Universitat Pompeu Fabra
ICREA
gemma.boleda@upf.edu

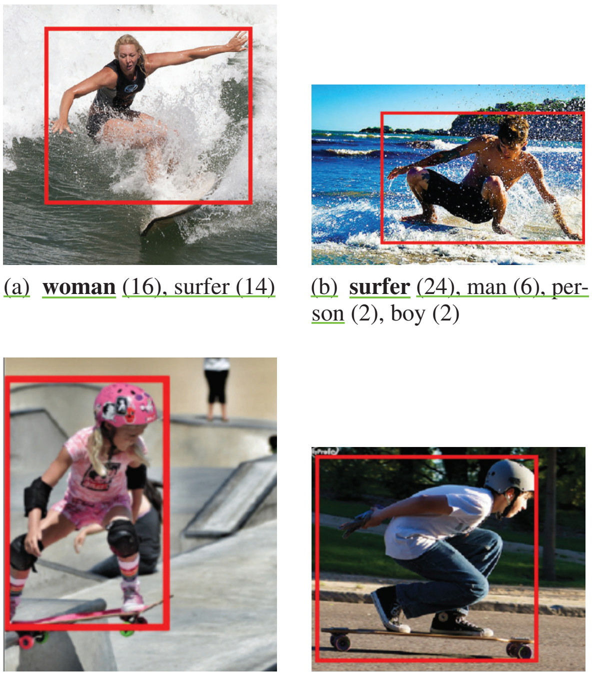
Abstract

Gender bias in Language and Vision datasets and models has the potential to perpetuate harmful stereotypes and discrimination. We analyze gender bias in two Language and Vision datasets. Consistent with prior work, we find that both datasets underrepresent women, which promotes their invisibilization. Moreover, we hypothesize and find that a bias affects human naming choices for people playing sports: speakers produce names indicating the sport (e.g. 'tennis player' or 'surfer') more often when it is a man or a boy participating in the sport than when it is a woman or a girl, with an average of 46% vs. 35% of sports-related names for each gender. A computational model trained on these naming data reproduces the bias. We argue that both the data and the model result in representational harm against women.

1 Introduction

Existing social biases and stereotypes against certain groups, such as women and racial minorities, are known to be reproduced by computational models (Caliskan et al., 2017; Bolukbasi et al., 2016; Hovy and Søgaard, 2015; Wang et al., 2021; Blodgett et al., 2020). This is primarily due to the fact that the datasets that the models are trained on are biased themselves, because, unless explicit steps are taken, datasets tend to mirror social biases (Torralba and Efros, 2011; Rudinger et al., 2017, 2018). Moreover, models often amplify biases, because they overrely on shallow patterns and lean towards majority labels (Ahmed et al., 2022; Deery and Bailey, 2022; Zhao et al., 2017).

Bias in AI has ethical implications because it can result in harm for the affected groups: both representational harm, with systems demeaning or ignoring them, and allocational harm, with systems allocating fewer resources or opportunities to them (Mitchell et al., 2021; Blodgett et al., 2020; Mehrabi et al., 2022b). For instance, the fact that the multimodal model VL-BERT (Su et al., 2019) often predicts that a woman carrying a briefcase is carrying a purse (Srinivasan and Bisk, 2022) constitutes representational harm, with working women not being recognized as such.


(a) **woman** (16), surfer (14)

(b) **surfer** (24), man (6), person (2), boy (2)

(c) **girl** (11), skateboarder (7), skater (6), child (6)

(d) **skater** (9), skateboarder (6), boy (6), kid (3)

Figure 1: Images of people playing sports from the ManyNames dataset, together with the names that human annotators produced and their counts.

In this paper, we focus on gender bias that causes representational harm for women, specifically for women in sports, in the area of Language and Vision (L&V). Previous work on gender bias in L&V has shown that bias often relates to the language component, causing models to override the specific visual information in classification decisions (Zhao et al., 2017; Goyal et al., 2017; Ramakrishnan et al., 2018) and vice versa (Hendricks et al., 2018; Bhargava and Forsyth, 2019), such as in the 'purse/briefcase' example above. We examine bias

Findings of the Association for Computational Linguistics: ACL 2023, pages 14093–14103
July 9-14, 2023 ©2023 Association for Computational Linguistics

that is present in the language, without examining its interaction with the visual component in the model.

Women in Western societies have traditionally been marginalized, excluded, and deterred from participating in sports or even physical activity, while men have been encouraged (Bell et al., 2008; Scheadler and Wagstaff, 2018; Vertinsky, 1994; Schaillée et al., 2021). Sport has long been stereotypically associated with masculinity, and females have been thought physically incapable of performing well in this area (Young, 1980). Participation in sports by women and girls has continued to increase since the 1960s in parallel with other social advances, however, rarely do women rise to managerial or coaching roles, even in women's teams (Schaillée et al., 2021). Women's sports are also underrepresented in the media, which has been linked to perpetuating the stereotype of the male athlete over the female (Schmidt, 2013).

We analyze the data in a Language and Vision dataset for object naming, ManyNames (Silberer et al., 2020a,b; more information in the next section). Figure 1 shows two example images in ManyNames together with the names elicited from subjects. How we name an object or entity is intimately linked to how we conceptualize it (Brown, 1958; LaTourrette and Waxman, 2020). We hypothesize that due to the social constraints discussed above, speakers produce a sports-related name such as 'surfer' less often when their referent is a woman, that is, they do not conceptualize female athletes as athletes. If the bias indeed exists in the speaker population, we expect it to be present in the naming data and propagate to computational models; we check both expectations. Before that, we examine whether there is an overall representational bias in ManyNames and its parent dataset, VisualGenome (Krishna et al., 2016), with women being underrepresented.

2 Underrepresentation of Women

Data and method ManyNames contains names for 25K images produced by English-speaking subjects in a free naming task, with an average of 31 names per image. ManyNames images are a subset of those in VisualGenome. The images were selected from seven previously defined domains, one of which was PEOPLE, based on a series of seed WordNet synsets. The authors put a cap on the number of images for a given synset (max. 500 instances for seeds with up to 800 objects in VisualGenome and up to 1k instances for seeds with more than 800 objects).

VisualGenome contains 108K images that are the intersection of images in the datasets YFCC100m (Thomee et al., 2016) and MS-COCO (Lin et al., 2015). The objects in each VisualGenome image were manually identified, labeled, and linked to their corresponding WordNet synset (VisualGenome provides many more annotations, but these are the ones of interest for the present study). Both VisualGenome and ManyNames employed crowd-sourced workers from Amazon Mechanical Turk (AMT) as annotators [1] for the images, with workers coming predominantly from the USA. Note that while the images in ManyNames are a subset of those in VisualGenome, the naming annotations were collected afresh for ManyNames.

The YFCC100m images were all those uploaded to Flickr between 2004 and 2014, published under a commercial or noncommercial license. MS-COCO contains all images from Flickr available at the time of the dataset construction that belonged to 91 predefined image categories (e.g., *horse*, *people*, and *laptop*). Images on Flickr are uploaded by Flickr users.

To check for representational bias, we extracted all objects in VisualGenome (image areas within a bounding box) with labels corresponding to the four most common gender-associated names: 'boy', 'girl', 'man', and 'woman'. For ManyNames, we used the same names and the full naming distribution.

Results The resulting gender distribution is shown in Table 1, together with the distribution in the world in 2020, as reported by the United Nations world population data (United Nations - Department of Economic and Social Affairs, 2022). Both datasets indeed underrepresent females, according to one-tailed z-tests comparing the percentage of females in each dataset with the global

[1] In VisualGenome, annotators identified (via bounding boxes) the objects in each image and provided descriptions such as 'a red mushroom with white spots'. The head nouns in these region descriptions were identified and matched with WordNet synsets in a semiautomatic fashion. In ManyNames, annotators were shown images where an object was highlighted with a bounding box, and they were asked to provide a single name for the object. Thirty-six different subjects provided names for each object, and the naming responses were cleaned and aggregated. We use the object synsets for the analysis in Section 2, and the naming distributions for the rest of the paper.

Names	VG	MN	World
woman	25.6	39.7	-
girl	7.0	7.9	-
man	59.1	37.9	-
boy	8.1	14.4	-
total female	**32.7**	**47.6**	**49.6**
total male	67.2	52.3	50.4

Table 1: Gender distribution in VisualGenome (VG), ManyNames (MN), and the world in 2020, in percentage.

percentage of females in the world (VisualGenome: z = -129.8, p < 0.001; ManyNames: z = -2.4, p = 0.02).[2] Note, however, that the bias against women is much larger in VisualGenome, with only 32.7% female entities compared with 49.6% in the world's population; in ManyNames, the percentage is only 2 points below the world population (47.6% vs 49.6%). Moreover, note that the bias in ManyNames stems from images of boys, with 14.4% of images vs 7.9% for girls. Recall from above that no specific action was taken in either ManyNames or VisualGenome regarding gender balance. The representational bias in ManyNames is smaller due to the cap on the synsets, aimed at obtaining a varied set of categories in general; the reduced gender bias is a side effect. Below, we discuss a further specific type of underrepresentation also found in ManyNames, that of women playing sports.

3 Sports-Related Bias

Next, we present our main analysis, namely gender bias related to sports as shown in human naming data. A secondary analysis concerns model behavior.

Sports-Related Bias in Humans: Methods The first author of this paper went through all topnames—that is, the name produced by the majority of the annotators for each image—in the ManyNames domain PEOPLE and selected those that related directly to gender ('boy', 'girl', 'man', and 'woman') or sport ('athlete', 'baseball player', 'basketball player', 'batter', 'catcher', 'goalie', 'pitcher', 'player', 'skateboarder', 'skater', 'skier', 'soccer player', 'snowboarder', 'surfer', 'tennis player', and 'umpire'). We refer to the former as 'taxonomic' and the latter as 'sports-related'. We selected all 1,776 images that have at least one taxonomic and one sports-related name in the responses, such that we could automatically determine both the gender of the person and the fact that they are playing a sport.[3]

To check for bias, we computed, for each image, the percentage of sports-related names associated with it relative to the total names, which include both taxonomic and sports-related names. For instance, in Figure 1, the person in panel (a) received 46.7% of sports-related names, while the person in panel (b) received 70.5%. We fitted a logistic regression model with the proportion of sports-related names as the outcome variable and fixed effects for the person's gender.

Sports-Related Bias in Humans: Results Table 2 summarizes the results of the logistic regression, supporting our hypothesis: when annotators see images of men playing sports, they are more likely to produce a name that explicitly mentions the sport being played and therefore are less likely to produce a taxonomic name compared with when they see images of women playing sports. Figure 2 qualitatively shows the difference between the genders. Note that there are also fewer images of women playing sports (527 vs. 1219), constituting only 30.2% of the pictures of people playing sports. These numbers compared with the real-world sports statistics constitute a further, more insidious instance of underrepresentation of women in L&V datasets: women in certain roles. This constitutes representational bias regarding women playing sports in ManyNames. Note that there are in general fewer women playing most sports in the Western population—especially due to dropouts (Bevan et al., 2021). However, the difference between the genders is not as large as in the dataset. For instance, in the US, the percentage of girls in college sports in 2019 was 43.9% *vs* 56.1% boys (NCAA, 2022), and in England, the percentage of women among the adults who participated in sporting activities in 2020 was 45% *vs* 55% men (Sport England, 2023).

[2] All statistical tests in this paper assume an alpha level of 0.05.

[3] Images with inconsistent gender in the response (e.g., where some subjects produced 'man' and others 'woman') were discarded.

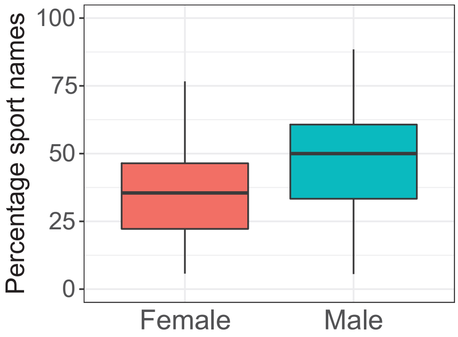


Figure 2: Distributions of percentages of sports-related names for female and male images in ManyNames.

Model; dep. var.: Prop sports-related names			
	Estimate	St. Error	p-value
Intercept	−0.62	0.02	p<0.001
GenderM	0.47	0.02	p<0.001
Descriptive statistics: Perc sports-related names			
	N	M	SD
female	527	34.9	16.9
male	1219	46.2	18.7

Table 2: Top: Fixed-effect estimates for the logistic regression model predicting the proportion of sports-related names based on image gender. Bottom: Descriptive statistics of the sample (N = number of images; M = mean; SD = standard deviation).

Sports-Related Bias in an L&V model We additionally analyze the behavior of a model trained to produce names on the ManyNames dataset (Lang, 2021). Based on the extensive literature on bias (Mehrabi et al., 2022a), we expect the model to reproduce the biases observed in the data; in contrast, it is unclear whether it will amplify it, as models vary with respect to this (Zhao et al., 2017; Fernando et al., 2021).

This model builds upon a ResNet101 architecture (He et al., 2015) pretrained on VisualGenome (Anderson et al., 2018) and is adapted to ManyNames names through an additional fine-tuning step. Relevantly to our purposes, the model is trained to reproduce the full naming distribution, outputting a probability distribution over all the names in the vocabulary. This is different from object classification in Computer Vision, which assigns a single class to each object (Deng et al., 2009; Ren et al., 2015; He et al., 2015). Lang (2021) used the train–dev–test partition of ManyNames established in Silberer et al. (2020b); we analyze the behavior of the model in the test set and, in particular, in the images that meet the criteria used in our second analysis above (N = 89, of which 34 female).

We normalize the probability weights output by the model, using only the names of interest (that is, discarding any weight the model places on names that are neither taxonomic nor sports-related before doing the normalization). To check whether the model reproduces the bias, we fit a logistic regression model, with the proportion of sports-related names as the dependent variable and fixed effects for the image gender. To check whether it amplifies the bias, we fit a mixed-effects logistic regression model that takes into account both the model and the human data (see Appendix for details). [4]

As shown in Table 3, according to the regression analysis, the L&V model indeed reproduces the naming bias found in the human data: for images depicting boys and men playing sports, it assigns significantly higher weights to sports-related names than for images depicting women or girls playing sports. We instead find no evidence of bias amplification (see Appendix); note, however, that the sample is small.

Model; dep. var.: Prop sports-related names			
	Estimate	St. Error	p-value
Intercept	−0.62	0.07	p<0.001
GenderM	0.63	0.08	p<0.001
Descriptive statistics: Perc sports-related names			
	N	M	SD
female	34	36.6	21.9
male	55	48.5	24.5

Table 3: Top: Fixed-effect estimates for the logistic regression model predicting the proportion of sports-related names based on image gender. Bottom: Descriptive statistics of the sample (N = number of images; M = mean; SD = standard deviation).

4 Discussion and Conclusions

We have identified pervasive biases against women in Language & Vision. Our main contribution is the individuation of a bias that characterizes human naming choices and therefore the naming data available for models. While we have focused on ManyNames, this issue is likely to affect other

[4] Models were fit in R using *glm* and *glmer* (Bates et al., 2015; R Core Team, 2021).

datasets containing names, such as widely used datasets for captioning (Young et al., 2014; Sharma et al., 2018; Lin et al., 2015) and referring expression generation (Kazemzadeh et al., 2014; Yu et al., 2016). The bias concerns the kind of name chosen for athletes, depending on the genre: people produce fewer sports-related names for females playing sports (average 35%) than for males playing sports (46%). As far as we know, this kind of bias has not been previously discussed, and it is, we argue, more implicit and thus difficult to identify than other kinds of bias that are more commonly discussed in the literature, such as unbalanced classes in datasets due to the underrepresentation of certain groups (which we also found and discuss below). We find the naming bias both in the human data of ManyNames and a model trained on these data. Thus, even women who do play sports (which are fewer to begin with, as mentioned in Section 3) are not conceptualized as such. This constitutes representational harm and contributes to limiting choices for women.

We also find that women are underrepresented in L&V, pronounced especially in VisualGenome, which has over twice as many images of males than females. Moreover, the proportion of males and females playing sports in ManyNames is skewed, with only 30.2% of the pictures of people playing sports depicting females. As mentioned above, the actual percentage of women and girls in sports in Anglo-Saxon societies is closer to 45%, according to recent data (NCAA, 2022; Sport England, 2023). The underrepresentation of social groups is harmful itself (Blodgett et al., 2020), as well as because models trained on unbalanced data can neglect crucial patterns relevant to those within the group (Wang et al., 2022). Given how the images were selected (see Section 2), the origin of the underrepresentation of women in these datasets must come from the kinds of images uploaded onto Flickr around the 2010s. This is in turn is likely rooted in the demographic characteristics of Flickr users during that period and to the fact that, in general, the internet has historically been heavily male-dominated (Morahan-Martin, 1998; O'Hare and Murdock, 2012). This is a serious concern, as most resources used in L&V (as well as computational linguistics and AI in general) come from the internet.

Our findings are in line with other research on gender bias, both in L&V and in NLP and AI more broadly, as discussed in the introduction. They also resonate with a study by the Pew Research Center (Lam et al., 2018) showing that results in Google Image Search underrepresent women in various jobs compared with their actual participation in those jobs in the USA according to the Bureau of Labor Statistics.

Ultimately, based on the findings of this paper, it can be concluded that as far as the datasets and model analyzed are concerned, when it comes to sports: A man with a tennis racket is a tennis player. A woman with a tennis racket is just a woman with a tennis racket.

Limitations and Future Work

Our findings about gender bias in the field of Language & Vision are based on two datasets, one task (Object Naming), one language (English), a mostly Western population (based on the origin of both the images and the annotators of VisualGenome and ManyNames), and one computational model. Moreover, in the third analysis, due to the characteristics of the test set of ManyNames, the sample size was small. Additionally, the bias around naming choices concerns the domain of sports only.

Regarding our most novel finding (bias in lexical choice), given the basic function of naming in language and the fact that Western English-speaking societies are not known to be more gender-biased than most non-Western and/or non-English-speaking societies, it is plausible that the identified bias extends to other L&V tasks such as image captioning, referring expression generation, or Visual Question Answering. Furthermore, given previous work on bias in our field, it is plausible that the identified bias in the model extends to other models. It is, however, not clear whether the bias will be amplified or simply reproduced. To probe whether, and to what extent, the identified biases indeed generalize, future work should tackle more tasks, languages, populations, domains, models, and data. Testing further models on the same naming data that we used is straightforward; checking for biases in some other tasks for English should be feasible at least to some extent, since some datasets provide multiple annotations per image (e.g., captions in MS-COCO). Instead, analyzing other languages and populations, such as nonbinary individuals, will in most cases require further data collection due to the scarcity of non-WEIRD data in our field.

In this study, gender was operationalized in a binary manner in order to most effectively investigate the stated hypothesis. Furthermore, there is a lack of nonbinary labels within the datasets used (5% of labels can be considered gender-neutral, i.e., "person, human, child"), and the resources required to reflect the reality of the gender landscape currently do not exist. This indicates a separate but related issue regarding a lack of representation of nonbinary individuals within vision datasets and how to conduct ethically inclusive studies on gender (Larson, 2017). However, addressing this is beyond the scope of the present research and remains an important direction for future work. Finally, this work solely concerns the identification of biases; further work should focus on how to deal with them in terms of data collection, curation, and modeling.

Ethics Statement

This research aims to highlight the extent to which gender biases may be present in Language & Vision, with an emphasis on the representation of females in sports. The findings of the study may have implications for the ways in which images of female athletes and sports figures are portrayed and treated within the datasets and by models. The results of the study may have the potential to contribute to the ongoing efforts to address gender bias in all areas of life, as we maintain that gender bias is an issue within Machine Learning, because it is an issue within human society. The research team is committed to using the findings of the study to foster dialogue and understanding of the issue and to advocate for equitable treatment of all groups throughout the production pipeline.

Acknowledgments

The authors thank the reviewers for their useful feedback. This research is an output of grant PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033, funded by the Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (Spain) and of grant agreement No. 715154 funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme. This paper reflects the authors' view only, and the funding agencies are not responsible for any use that may be made of the information it contains.

References

Md. Arshad Ahmed, Madhura Chatterjee, Pankaj Dadure, and Partha Pakray. 2022. The role of biased data in computerized gender discrimination. GE@ICSE '22, page 6–11, New York, NY, USA. Association for Computing Machinery.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Richard C. Bell, Ed D., and J. D. 2008. A history of women in sport prior to title ix. *The Sport Journal*, 10.

Nadia Bevan, Claire Drummond, Liz Abery, Sam Elliott, Jamie Lee Pennesi, Ivanka Prichard, Lucy K. Lewis, and Murray Drummond. 2021. More opportunities, same challenges: adolescent girls in sports that are traditionally constructed as masculine. *Sport, Education and Society*, 26(6):592–605.

Shruti Bhargava and David Forsyth. 2019. Exposing and correcting the gender bias in image captioning datasets and models.

Su lin2015 Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational lin2015guistics*, pages 5454–5476, Onlin2015e. Association for Computational lin2015guistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

R. Ben Brown. 1958. How shall a thing be called. *Psychological review*, 65 1:14–21.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Oisín Deery and Katherine Bailey. 2022. The bias dilemma. *Feminist Philosophy Quarterly*, 8(3).

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Martínez-Plumed Fernando, Ferri Cèsar, Nieves David, and Hernández-Orallo José. 2021. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7):3217–3258.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. *European Conference on Computer Vision*, pages 771–787.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.

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, Doha, Qatar. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Onyi Lam, Stefan Wojcik, Brian Broderick, and Adam Hughes. 2018. Gender and jobs in online image searches. Pew Research Center.

Fabian Lang. 2021. What affects object naming: Reconstructing human naming of objects in complex images. Master's thesis, Universität Stuttgart, Germany.

Brian N Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. Association for Computational Linguistics.

Alexander LaTourrette and Sandra Waxman. 2020. Naming guides how 12-month-old infants encode and remember objects. *Proceedings of the National Academy of Sciences*, 117:202006608.

T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. 2015. Microsoft COCO: Common object in context.

Ninareh Mehrabi, Ahmad Beirami, Fred Morstatter, and Aram Galstyan. 2022a. Robust conversational agents against imperceptible toxicity triggers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2831–2847, Seattle, United States. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022b. A survey on bias and fairness in machine learning.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

Janet Morahan-Martin. 1998. The gender gap in internet use: Why men use the internet more than women—a literature review. *CyberPsychology & Behavior*, 1(1):3–10.

NCAA. 2022. The state of women in college sports.

Neil O'Hare and Vanessa Murdock. 2012. Gender-based models of location from flickr. In *Proceedings of the ACM Multimedia 2012 Workshop on Geotagging and Its Applications in Multimedia*, GeoMM '12, page 33–38, New York, NY, USA. Association for Computing Machinery.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming language priors in visual question answering with adversarial regularization.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. *Association for Computational Linguistics*, Proceedings of the First ACL Workshop on Ethics in Natural Language Processing:74–79.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 8–14. Association for Computational Linguistics.

Hebe Schaillée, Inge Derom, Oskar Solenes, Solveig Straume, Beth Burgess, Vanessa Jones, and Gillian Renfree. 2021. Gender inequality in sport: perceptions and experiences of generation z. 26:1011–1025.

Travis Scheadler and Audrey Wagstaff. 2018. Exporsure to women's sports: Changing attitudes toward female athletes. *The Sport Journal*, 20.

Hans C. Schmidt. 2013. Women, sports, and journalism: Examining the limited role of women in student newspaper sports reporting. *Communication & Sport*, 1(3):246–268.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational lin2015guistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational lin2015guistics.

Carina Silberer, Sina Zarrieß, and Gemma Boleda. 2020a. Object naming in language and vision: A survey and a new dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5792–5801, Marseille, France. European Language Resources Association.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020b. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sport England. 2023. Sport england.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85, Seattle, Washington. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations.

B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. J. Li. 2016. YFCC100M: The new data in multimedia research. 59(2):64–73.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. *CVPR 2011*, pages 1521–1528.

United Nations - Department of Economic and Social Affairs. 2022. Population division.

Patricia A. Vertinsky. 1994. Gender relations, women's history and sport history: A decade of changing enquiry, 1983-1993. *Journal of Sport History*, 21(1):1–24.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Peifeng Wang, Filip Ilievski, Muhao Chen, and Xiang Ren. 2021. Do language models perform generalizable commonsense inference? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3681–3688, Online. Association for Computational Linguistics.

Iris Marion Young. 1980. Throwing like a girl: A phenomenology of feminine body comportment motility and spatiality*. *Human Studies*, 3:137–156.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modelin2015g context in referring expressions.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

A Bias Amplification

To check whether the model amplifies the bias, we considered, for the same 89 images, the probabilities assigned to taxonomic and sports-related names by the human annotators and the model and fitted a mixed-effects logistic regression model, with the proportion of sports-related names as the outcome variable and fixed effects for the image gender, the prediction type (either *human* or *model*), and the interaction between image gender and prediction type. We set a random intercept for each image and a random slope for each type. We coded the prediction type as 'treatment' so that the *human* prediction would be our baseline.

The results are summarized in Table 4. The regression shows that for the 89 images included in this analysis, humans use more sports-related names when the image gender is male—in line with our previous findings. However, the results concerning a possible bias amplification are inconclusive: the estimates for the prediction type and for the interaction between type and gender are not significant. Analyses of larger data samples may be needed to shed further light on this topic.

Dependent variable: Prop sports-related Names			
	Estimate	St. Error	p-value
Intercept	-1.05	0.15	p<0.001
GenderM	0.99	0.19	p<0.01
TypeM	0.25	0.20	p<1
GeM:TyM	-0.26	0.263	p<1
Descr. stats., model: Perc sports-related names			
	N	M	SD
female	34	36.6	21.9
male	55	48.5	24.5
Descr. stats., human: Perc sports-related names			
	N	M	SD
female	34	28.2	18.3
male	55	49	18.3

Table 4: Fixed-effect estimates for mixed-effects logistic regression model predicting the proportion of sports-related names based on image gender and type of prediction (human *vs* model—with human-treated as baseline).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☒ A2. Did you discuss any potential risks of your work?
We do not think this is relevant to our submission.

☒ A3. Do the abstract and introduction summarize the paper's main claims?
The abstract does, the introduction no because it's a short paper and we distribute the claims between the introduction and the conclusion.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

All sections.

☑ B1. Did you cite the creators of artifacts you used?
Introduction.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We will add this to the final version, if accepted.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
It follows from the content of the paper.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We will add this to the final version, if accepted.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Introduction, sections 3 and 5, limitations and future work.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.