Vision-Language Models Performing Zero-Shot Tasks Exhibit Disparities Between Gender Groups

Melissa Hall Laura Gustafson Aaron Adcock Ishan Misra Candace Ross
Meta AI
melissahall@meta.com, ccross@meta.com

Abstract

*We explore the extent to which zero-shot vision-language models exhibit gender bias for different vision tasks. Vision models traditionally required task-specific labels for representing concepts, as well as finetuning; zero-shot models like CLIP instead perform tasks with an open-vocabulary, meaning they do not need a fixed set of labels, by using text embeddings to represent concepts. With these capabilities in mind, we ask: Do vision-language models exhibit gender bias when performing zero-shot image classification, object detection and semantic segmentation? We evaluate different vision-language models with multiple datasets across a set of concepts and find (i) all models evaluated show distinct performance differences when identifying concepts based on the gender of the person co-occurring in the image (ii) model calibration (*i.e.*, the relationship between accuracy and confidence) also differs distinctly by gender, even when evaluating on similar representations of concepts and (iii) these observed disparities align with existing gender biases in word embeddings from language models. These findings suggest that, while language greatly expands the capability of vision tasks, it can contribute to propagating social biases in zero-shot settings.*

1. Introduction

Natural language has greatly expanded the capabilities of vision models during inference, going from fixed vocabularies of visual concepts to essentially limitless concepts. Vision-language models, such as CLIP [20] and ALIGN [12], are a powerful means for representation learning of concepts. These models have impressive zero-shot image recognition capabilities wherein, at test time, the language embeddings of new visual classes can serve as a classifier. While such a broad range of recognition abilities is convenient, it also makes these models harder to analyze from a fairness perspective as the model's recognition vocabulary is not fixed and is infinitely large.

Prior works focus on measuring biases in multimodal word embeddings [21, 24] and language and vision models separately [5, 16, 3, 8]. Other works measure differential performance of multi-modal models for a small specific vocabulary, often adversarially tuned [9], and find that models contain social biases in perpetuating harmful stereotypes around criminality and dehumanization via disproportionate associations between demographic groups [1]. Thus, while works extensively study models with a fixed vocabulary and set of tasks, they do not inspect the performance of language-vision models in zero-shot settings and compare with upstream biases.

We measure and explore the gender bias in zero-shot, multi-label image classification by probing CLIP as well as two downstream tasks of object detection using Detic [26] and semantic segmentation LSeg [14]. Probing these three zero-shot vision-language models for gender bias, our contributions are as follows:

1. We show that zero-shot vision-language models show gender-based performance disparities for different visual concepts. This means that, for a given concept, the model will perform better when the concept co-occurs with one gender as opposed to another.
2. For object detection and segmentation models, we find that calibration between model performance and confidence also differs by gender across concepts.
3. Lastly, we find that the biases in word embeddings from word2vec parallel the biases we find in the zero-shot vision-language models.

2. Framework

Our analysis focuses on evaluating zero-shot classification, detection, and segmentation models. In particular, we investigate CLIP$_{ViT-B/32}$ [20, 7], a contrastive model that, given an image and corresponding text, outputs the cosine similarity between the two inputs using separate language and vision encoders. Due to its design, CLIP can perform zero-shot image classification by using an object class as the text input. We also study two models that utilize CLIP

for zero-shot detection, Detic [26], and segmentation, LSeg [14]. Our models are evaluated in the zero-shot setting.

To allow for a side-by-side of comparison of performance disparities between the models with a shared metric, we adapt the detection and segmentation models to support multi-label image classification. Assume a set of images I and a set of object classes C[1], where each image has a set of ground-truth object classes present in the image and a single associated gender label. For a given image $i \in I$, a (multi-label) image classifier outputs a set of object classes that are predicted to be contained in the image; object detection outputs N bounding boxes, where each bounding box is associated with a single label $c \in C$; and semantic segmentation produces a single label $c \in C$ for each pixel in the image. We use Detic's bounding box object predictions and LSeg's pixel-by-pixel classifications as binary indicators of the model's recognition of the concept in the image.

For each concept $c \in C$, we explore how these models differ during inference when the gender of the people that are depicted in images changes. Specifically, we focus on images containing men versus images containing women. We use the Visual Genome dataset, which contains 108k images, where each image is annotated with a set of bounding boxes and one object label per box. The labels correspond to synsets from WordNet [18], where each synset is a node representing a singular concept in a tree of nodes. Visual Genome contains human-annotated synsets tagging objects (*e.g.* labels such as baseball bat and dress) and people (*e.g.* labels such as bride, person, doctor). We use the synset labels to determine the set of objects in the image and gender of the people present. To designate gender groups, we create two sets of labels – one corresponding to concepts referring to women (*e.g.* mother, wife) and the other corresponding to concepts referring to men (*e.g.* son, groom). See Table 1 in the Appendix for the full synset mapping. After getting the annotated gender labels, we retain only the images that are annotated with a single gender label. To increase reliability of our measurements, we use only object labels that occur in at least 50 images for each gender group. After filtering, we have 25,215 images and 408 object classes for Visual Genome.

3. Evaluation Setup & Findings

We adapt existing metrics to determine whether three vision-and-language models, CLIP, Detic and LSeg, have disparate gender-based performance for image classification between concepts, and to what extent these biases are similar to those from language models. We use images from the Visual Genome with filtered sets of images of men $\mathcal{D}_{men}$ and woman $\mathcal{D}_{women}$ as described in Section 2. We use publicly available pretrained checkpoints for all models and set the Detic minimum score for a predicted bounding box to be retained as 0.1 See Appendix A.2 for similar findings evaluated on the MS-COCO dataset [15].

[1]Note that we use object classes and concepts interchangeably.

3.1. Experiment 1: Disparity in average precision

Setup We ask whether a model has similar performance for image classification for a given concept when evaluating images in $\mathcal{D}_{men}$ and $\mathcal{D}_{women}$. To determine whether each model has differential performance across genders, we use the difference of average precision (AP) between $\mathcal{D}_{men}$ and $\mathcal{D}_{women}$ for every concept. The average precision is the weighted mean of model precision across concepts.We use AP as it is a popular metric for vision tasks and it accounts for a variable number of objects between images.

Results We find that *all models show differential outcomes by gender*, performing disparately between gender groups $\mathcal{D}_{men}$ and $\mathcal{D}_{women}$ for many concepts. Figure 1(a) shows the top differences in average precision (AP) for synsets that co-occur with man- and woman-annotated images for CLIP, Detic, and LSeg. Positive AP differences for a concept indicate better outcomes for images in $\mathcal{D}_{men}$ (*e.g.* the model performs better for images containing men over images containing women for the concept `necktie`) and negative AP differences signal a better measurement for images in $\mathcal{D}_{women}$. The direction of the AP differences is consistent among all three models for many objects, indicating shared fairness concerns among each of them.

Furthermore, it is common to report the mean average precision across all concepts as a single, summary statistic of model performance [26]. Figure 2 demonstrates that the aggregation of AP across all concepts can mask these disparities and lend a false assurance of model consistency across demographic groups.

3.2. Experiment 2: Disparity in calibration

Setup We next explore how models treat different groups using *calibration*: If a model's calibration is similar between groups, it means the model assigns similar probability scores for samples that have the same expected likelihood of containing the concept, regardless of which group is depicted in the image. As an example, suppose we have two images - one containing a man, the other containing a woman - with the same expected likelihood that they contain the concept `necktie`. If the classification model assigns a lower confidence score to the image with a woman than the expected likelihood for `necktie`, while assigning a higher-than-expected probability to the image with a man, then the model is displaying a disparity in calibration.

Following previous work [10, 11], we study model calibration using the expected calibration error (ECE) [19], which is the absolute difference between the model's confidence and accuracy. Larger values of ECE mean greater

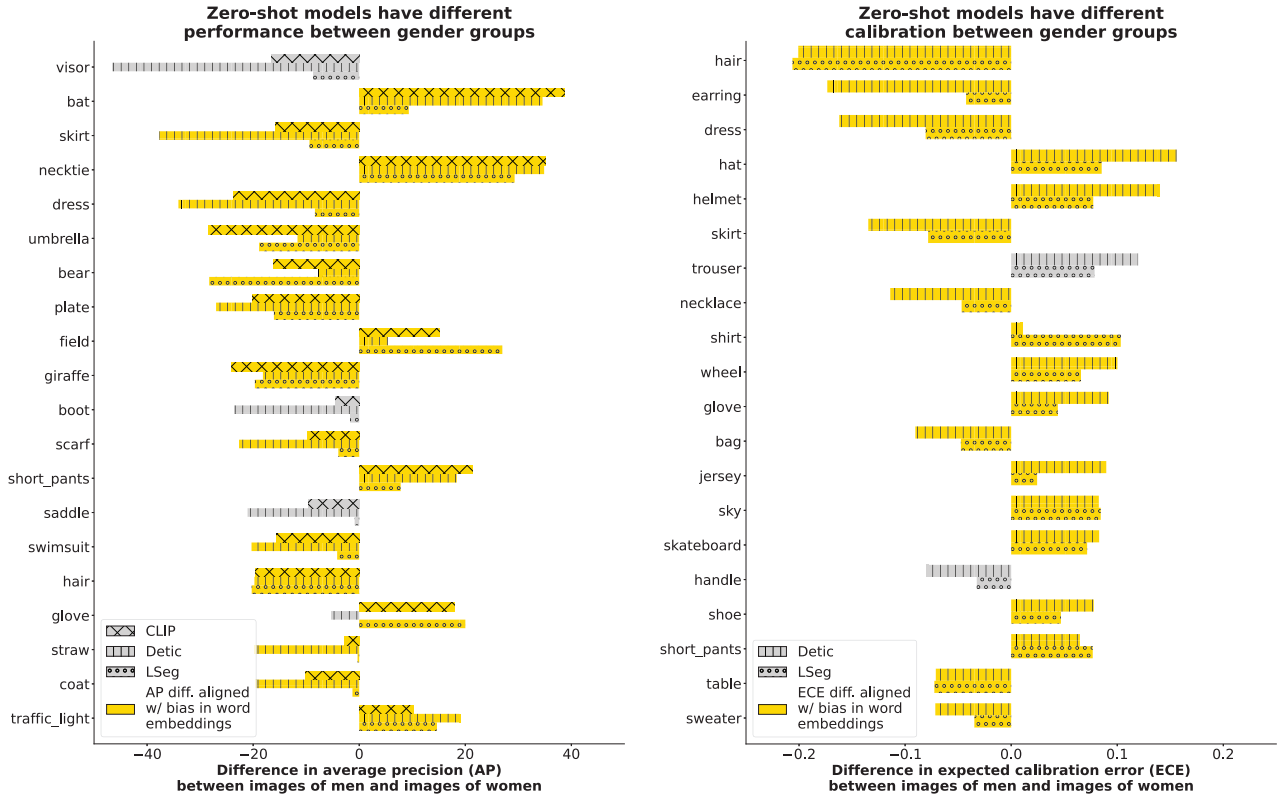

Figure 1. (a) The average precision (AP) gender disparity by concept for the Visual Genome, where positive values indicate better performance for images of men, $\mathcal{D}_{men}$, and negative values are better performance for images of women, $\mathcal{D}_{women}$. (b) We measure expected calibration error (ECE), which is the absolute difference between model confidence and accuracy, for each concept. For both figures, the bars colored yellow bars show AP/ECE disparities that also align with the social biases in word2vec embeddings.

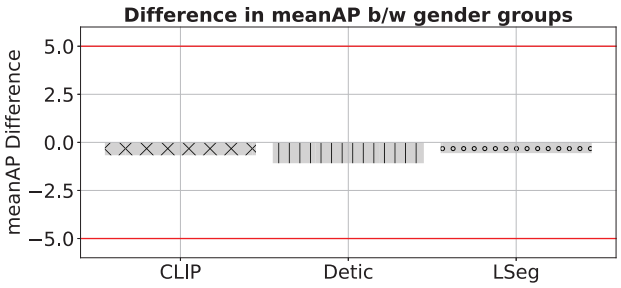

Figure 2. For CLIP, Detic, and LSeg evaluated on Visual Genome, the difference in meanAP between the annotated gender groups masks significant per-concept disparities observed in Figure 1.

model miscalibration. We evaluate Detic and LSeg, excluding CLIP because its multi-label classification setting does not produce probabilities (see Appendix A.3).

Results We find that *Detic and LSeg both show differential calibration for gender*. For each gender we computed the ECE for each concept and took the difference between $\text{ECE}_{\mathcal{D}_{men}}$ and $\text{ECE}_{\mathcal{D}_{women}}$; a positive ECE difference means the model is more calibrated for the given concept for images of men and a negative ECE means the model is more calibrated for images of woman. Figure 1(b) shows that many concepts have a large difference in ECE between groups.

3.3. Experiment 3: Relationship between bias in word embeddings and bias in zero-shot vision-language models

Setup For the final experiment, we explore whether the observed disparities in zero-shot models correlate with disparities found in word embeddings. Gender bias in word embeddings has been explored using the geometry of the embedding space [4] and average cosine similarities between sets of gendered terms and sets of stereotypical, non-gendered terms [5, 16, 23]. Accordingly, we extract embeddings for each concept using word2vec [17] trained on Google News 300M and compute the cosine similarity between each concept and the gendered terms in the embedding space following the method defined in Appendix A.4.

Results We further observe that *the disparities found in Experiments 1 and 2 for zero-shot vision-language models*

are aligned with those in word embeddings, as indicated by the cosine similarities. The bars in Figure 1 are colored yellow when the difference in cosine similarities between the word2vec text embeddings for the synset and gender terms are aligned with the discrepancies in AP and ECE. This suggests that language can contribute to social biases in vision-language models, particularly in zero-shot settings that do not perform any further finetuning.

3.4. Root Cause Analyses

We analyzed concepts in Visual Genome with disparities across gender for the three models and found several potential root causes.

First, definitions of concepts can differ between groups. For example, we observe images of halter swimsuits (which tie at the neck) labeled as "necktie.n.01", in addition to the more common necktie traditionally worn with dress-suits. Because women tend to wear such swimsuits more than men, a model that recognizes "necktie" in the more common sense may perform less well for women. In addition, images can vary in salience between gender groups, as the two groups depicted in the photos tend to co-occur with the concepts differently. For example, hair may be more prominent in images of women, who tend to have longer hair, than images of men. This variation in salience of a given concept between groups likely affects models' predictive performance for the group.

This suggests that our findings surface disparities that have real-world implications and can inform potential mitigation strategies.

4. Discussion

Natural language supervision has greatly expanded the capabilities of vision models. Many of these models are able to perform zero-shot image classification, object detection and semantic segmentation on an open-vocabulary. We probed three models – CLIP, Detic and LSeg – to see whether there were gender-based disparities in their performance and treatment for different groups and to see whether these disparities, if any, paralleled those found in word embeddings from language models.

We find that all of these zero-shot models perform differently for many concepts based on the gender of the co-occurring person in the image across multiple datasets. We also find that the relationship between model confidence and accuracy differs by gender for many concepts. These results show the importance of considering model fairness when using an open vocabulary in zero-shot settings. They also show that only measuring model performance without disaggregating by concept can mask model bias.

We hope this work paves the way for future investigation of these concerns, such as isolating how biases in language models can be a contributing factor to social biases in zero-shot setting or studying potential mitigations. While our evaluation serves as an initial method for auditing zero-shot vision models for demgoraphic disparities, it also suggests a need for future studies into how choices such as dataset, group definition, and metric impact disparity analyses of zero-shot and multi-label classification systems and for alignment on insightful and reliable evaluation protocols for these modern settings.

4.1. Limitations

We note that there are multiple issues with using annotations to approximate gender that plague most vision datasets used for disparity evaluations. The binarization of gender using synsets is reductive and excludes other genders not captured within WordNet. Also, this approach relies on annotators' inherent perception of gender and can lead to the misgendering of individuals depicted. Reliance on static, external annotations based on visual representations is inherently misaligned with an inclusive operationalization of gender [6]. This is a particularly prevalent issue in image and multi-modal datasets [22]. People depicted should be given the agency to optionally share and update their gender information throughout the dataset's lifespan.

Furthermore, while we do our best to perform rigorous and robust measurements, each decision made in a disaggregated evaluation of model performance may affect the observed findings [2]. As an example of one evaluation decision, we adapted the detection and segmentation tasks performed by Detic and LSeg, respectively, as multi-label classification tasks to enable a comparison of the three models between shared metrics and datasets. Other factors such as co-occurrence with other objects and variations in object size or image quality may also affect findings.

4.2. Broader Impacts

The importance of understanding societal effects of the use of representations from natural language to enable vision tasks only increases as such practices become more ubiquitous. Model fairness can be defined in many ways and the method of evaluation can reveal different patterns of disparities [13]. Our study highlights one of several ways to evaluate vision-language systems.

While understanding and minimizing observed disparities in model performance is a valuable goal in itself, it may be insufficient for ensuring that machine learning predictions are unbiased and fair. Optimizing models to reduce these disparities requires tradeoffs between other fairness guarantees and performance measures.

Acknowledgments

We thank Adina Williams, Laurens van der Maaten, Nicolas Usunier, Hervé Jégou, Bobbie Chern, Rebecca Qian, and Ranjay Krishna for their feedback on this work.

References

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating CLIP: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 1

[2] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna M. Wallach. Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs. *CoRR*, abs/2103.06076, 2021. 4

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. 1

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016. 3

[5] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 1, 3, 6

[6] Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of" gender" in nlp bias research. *arXiv preprint arXiv:2205.02526*, 2022. 4

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[8] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022. 1

[9] Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for systematic assessments of visual feature extractors, 2022. 1

[10] C. Guo, G. Pleiss, Y. Sun, and K.Q. Weinberger. On calibration of modern neural networks. pages 1321–1330, 2017. 2

[11] Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. A systematic study of bias amplification. *CoRR*, abs/2201.11706, 2022. 2

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[13] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016. 4

[14] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven Semantic Segmentation. In *International Conference on Learning Representations*, 2022. 1, 2

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[16] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1, 3

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 3

[18] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995. 2

[19] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. 2015. 2

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[21] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, 2021. 1

[22] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), may 2020. 4

[23] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[24] Jialu Wang, Yang Liu, and Xin Eric Wang. Assessing multilingual fairness in pre-trained multimodal representations. *arXiv preprint arXiv:2106.06683*, 2021. 1

[25] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *CoRR*, abs/1707.09457, 2017. 6

[26] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting Twenty-thousand Classes using Image-level Supervision. 2022. 1, 2

A. Appendix

A.1. Annotations Used for Group Assignment

The mapping of synset and caption terms used for defining groups in the Visual Genome and COCO datasets are shown in Table 1. This list is adapted from [5, 25].

A.2. Measurements with COCO

Setup In addition to the Visual Genome dataset, we also show results for MS-COCO [15], a dataset of 123k images containing a set of bounding boxes for 80 object categories and 5 captions per image. Because one of our models of evaluation uses COCO for training, we use the COCO 2017 validation set.

Because MS-COCO does not include synsets for each image, we use the captions to extract the perceived gender(s) in image following previous work [25]. Following our approach for Visual Genome, we create a list of gendered terms (see Appendix A.1 for full list) and keep only those images with a single gender reference across the captions. We use the train set for filtering objects that are not in at least 100 images. This leaves us with 1412 total images and 76 object classes.

We then map the gender-related COCO objects to the Visual Genome synsets in order to compare differential model performance on similar objects between two different datasets. For each COCO object, we find a Visual Genome synset with the same name or similar name. When we have the choice between multiple synsets or multiple synset definitions, we use the Visual Genome synset definition that most closely aligns with the object's representation in COCO based on visual inspection.

To summarize our results, objects are highlighted according to whether the AP differences are practically significant and indicate whether the trend in object performance is consistent between datasets:

- **Green**: The AP-differences for both Visual Genome and COCO are in the same direction and have magnitudes greater than 5, meaning that the performance is practically significantly higher for same gender-annotated group for both datasets.

- **Red**: The AP-differences for the two datasets are in opposite directions and have magnitudes greater than 5, indicating that the disparity in performance is significant yet not consistent between the two datasets.

- **Yellow**: At most one AP-difference between the two datasets is greater than 5. This implies that, while the model could favor one group in a dataset and vice versa for the alternative dataset, this difference is not practically significant for both datasets.

In short, green represents alignment in disparity concerns between datasets and red represents misalignment. We omit the objects that are not included in both datasets.

Results Figure 3 demonstrates the overlap in objects between Visual Genome and COCO for which CLIP and Detic have the highest disparity in AP. We find that there is a significant number of objects with consistent concerns of performance disparity between annotated binary gender groups across both datasets for both models (as indicated in green).

Specifically, in Figure 4 we see that there are 20 and 15 objects of practically significant, directionally similar performance concerns between the two datasets for CLIP and Detic, respectively. This provides additional evidence that the concerns observed in Visual Genome are not dataset specific and may be pervasive for the given objects even among different distributions of representation.

A.3. Why Experiment 2 Does Not Include CLIP

For Experiment 2 as described in Section 3.2, we measure the disparity in the expected calibration error (ECE) for a given concept across the genders we are evaluating. We only measure and report ECE for Detic and LSeg. If we were in a single-label classification setting with N object classes, we could take the output of CLIP (cosine similarities ranging from -1 to 1 between the N classes and the image) and take the softmax to produce probabilities. The probabilities correlate to model confidence and can be used to measure the ECE (where both confidences and accuracies range from 0 to 1). In the multi-label setting, we instead have the raw logits and cannot compute the softmax because multiple classes can be present in the image. We will consider different approaches in the future to measure calibration in the multi-label setting for CLIP.

A.4. Cosine Similarity Measurements

To perform cosine similarity measurements in Experiment 3, we first define a set of embeddings corresponding to each gender group following prior work [5], where each gender group contains multiple related terms. For example, the "woman" group set contains terms including "female", "woman", "girl", etc. We then define a mapping of embeddings corresponding to each synset: We use the synset itself for all concepts where the synset is a canonical term of reference for that concept and the synset consists of only one word (e.g. "refrigerator.n.01"). When the concept consists of two words (e.g. "electric_refrigerator.n.01"), we average the embeddings between the two words. When the synset is ambiguous for that concept and may be confused with other synsets related to the same concept, we select a modifier based on the synset definition (e.g. "knob.n.02" becomes "knob handle" and "helmet.n.02" becomes "protective helmet"). For each gender group, we average the

Visual Genome

man-related terms	*woman-related terms*
man.n.01,	maid.n.02,
male_child.n.01,	woman.n.01,
guy.n.01,	girl.n.01,
male.n.01,	lady.n.01,
groom.n.01,	female.n.01,
husband.n.01,	mother.n.01,
grandfather.n.01,	lass.n.01,
son.n.01,	ma.n.01,
boyfriend.n.01,	widow.n.01,
brother.n.01,	bride.n.01,
grandson.n.01,	daughter.n.01,
groomsman.n.01,	grandma.n.01,
ex-husband.n.01,	granddaughter.n.01,
uncle.n.01,	bridesmaid.n.01,
godfather.n.01	girlfriend.n.01,
	sister.n.01,
	wife.n.01,
	female_child.n.01,
	white_woman.n.01,
	dame.n.01,
	matriarch.n.01,
	mother_figure.n.01,
	dame.n.02,
	great-aunt.n.01,
	donna.n.01

MS-COCO

man-related terms	*woman-related terms*
man,	woman,
mans,	womans,
men,	women,
boy,	girl,
boys,	girls,
father,	lady,
fathers,	ladies,
son,	mother,
sons,	mothers,
he,	daughter,
his,	daughters,
him	she,
	her,
	hers

Table 1. The synsets for Visual Genome (left) and words from captions for MS-COCO (right) that we use to determine group membership for gender for the images in the datasets. We exclude the images annotated with synsets or captions that correspond the concept `person` or `people` as well as images that correspond to both man-related and woman-related terms.

cosine similarity between the synset embeddings and each gender term. We then use the difference in the cosine similarites between the "man" and "woman" gender groups as an indicator of the social bias for that concept: When the difference in average cosine similarity is positive, the concept is more aligned with "man" terms than "woman" terms and vice versa for negative differences in average cosine similarity.

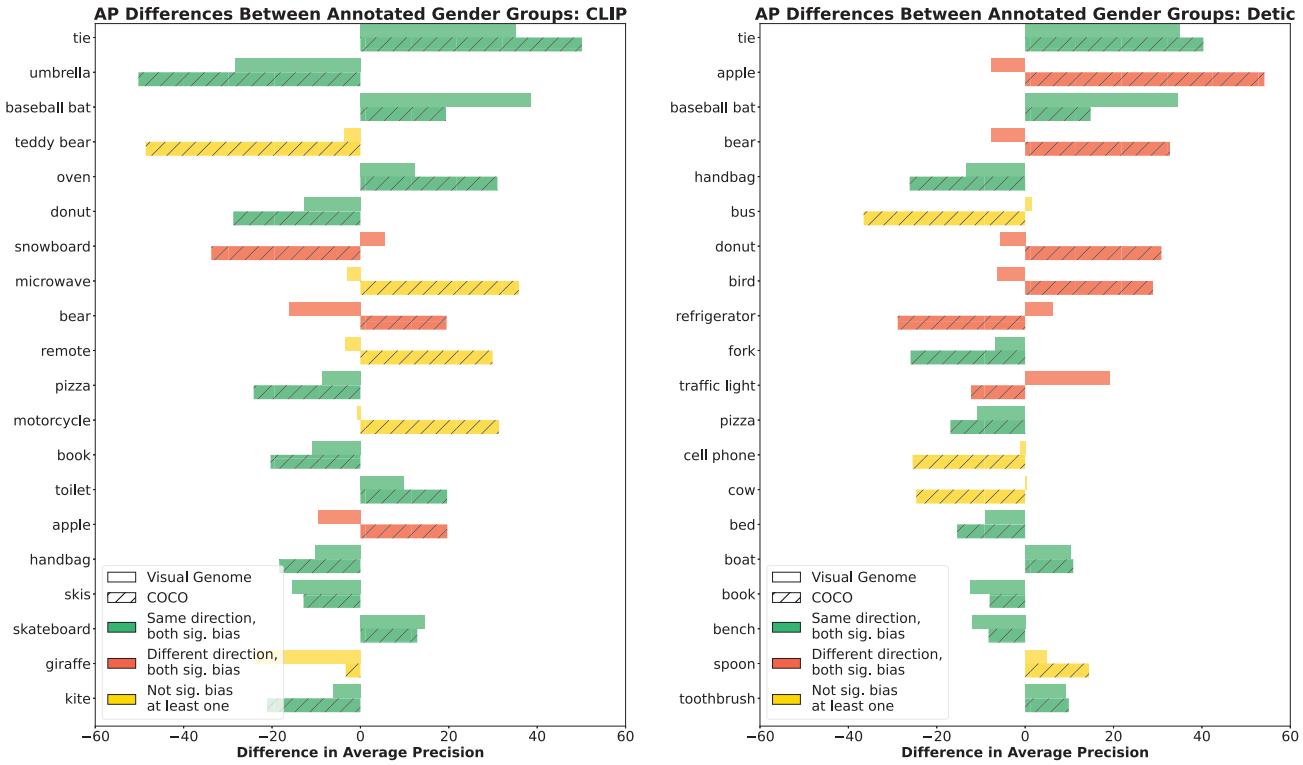

Figure 3. Objects with the highest-disparity in AP for CLIP (**left**) and Detic (**right**), evaluated on both Visual Genome and COCO.

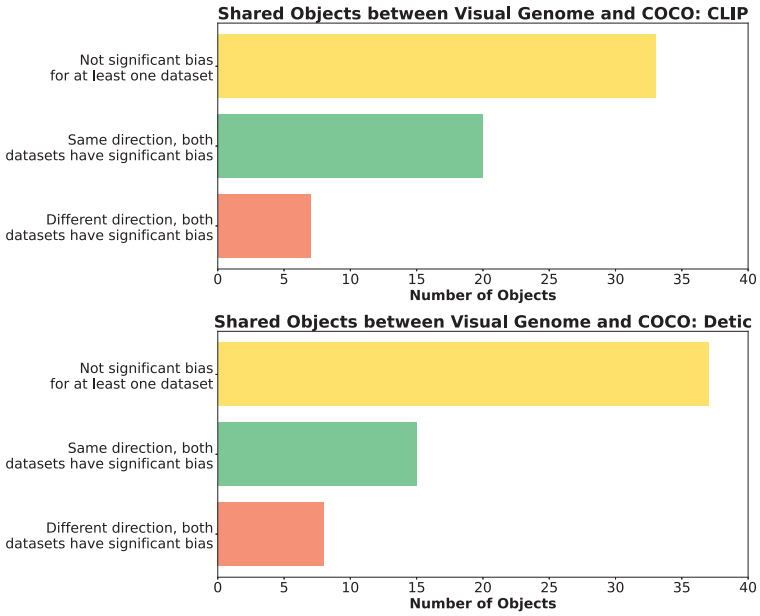

Figure 4. Many of objects shared between Visual Genome objects and COCO have a practically significant, directionally similar difference in AP for both CLIP (**left**) and Detic.