EVALUATING MODEL BIAS REQUIRES CHARACTERIZING ITS MISTAKES

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to benchmark model performance in the face of spurious correlations is important to both build better predictors and increase confidence that models are operating as intended. We demonstrate that *characterizing*, as opposed to simply quantifying, model mistakes across subgroups is pivotal to properly reflect model biases which are ignored by standard metrics such as accuracy gap. Inspired by the hypothesis testing framework, we introduce SKEWSIZE, a principled and flexible metric that captures bias from mistakes in a model's predictions and can be used in multi-class settings or generalised to the open vocabulary setting of generative foundation models. SKEWSIZE is an aggregation of the *effect size* of the interaction between two categorical variables: the spurious variable representing the bias attribute the model's prediction. We demonstrate the utility of SKEWSIZE in multiple settings including: standard vision models and multimodal foundation models from the BLIP-2 family. In each case, the proposed SKEWSIZE is able to highlight biases not captured by other metrics, while also providing insights on the impact of recently proposed techniques, such as instruction tuning.

1 INTRODUCTION

Machine learning systems can capture unintended biases (Dixon et al., 2018) by relying on correlations in their training data that may be spurious (i.e. a finite sample artifact), undesirable and/or that might vary across environments. Models of all scales are vulnerable to this failure mode, including recent, large-scale models (Weidinger et al., 2022; Birhane et al., 2023; Luccioni et al., 2023). To evaluate unintended biases in model outputs, existing metrics divide the population into subgroups and aggregate the e.g. correct and incorrect outputs across those subgroups as in Sagawa et al. (2019). However, this process consider as equivalent all responses deemed to be incorrect, obscuring important information regarding a model's bias characteristics, especially in the context of large or intractable output spaces. We introduce an example of such situations in Appendix A where three accuracy-based metrics fail to capture biases that appear in the mistakes of two models.

To measure this type of bias, we introduce SKEWSIZE, which considers *how different* the distribution of predictions are across subgroups. In our motivating example, SKEWSIZE is able to capture the different types of biases. We propose to formulate the problem of estimating bias for classification models through the lens of hypothesis testing. We draw inspiration from tests of association between the confounding, spurious factor and the model's prediction, and propose to re-purpose a measure of *effect size* for such tests. We compute effect sizes of this association for each ground-truth class: e.g., given images of doctors, we can estimate the effect size of the association between gender and predicted occupation. We show this approach yields a fine-grained and interpretable assessment of model bias, exposing the most affected classes, as opposed to accuracy-based or fairness metrics. Finally, we propose to aggregate effect sizes across classes using a measure of the skewness of the effect size distribution per class to obtain at a scalar metric which can be used to compare models.

We validate the metric and investigate its utility in multiple settings, including foundation vision-and-language models (VLMs) (BLIP-2, Li et al. (2023)) that have an intractable[1] output space in two settings: gender vs. occupation and practiced sport. Our main contributions are summarized as:

[1]This refers to the setting where the label space is given by all the possible outputs of a language model.

1. We demonstrate limitations of current metrics for quantifying bias, specifically that they fail to capture bias manifested in *how the model makes mistakes*.
2. We propose SKEWSIZE, a metric for evaluating bias in discriminative models inspired by hypothesis tests of contingency tables.
3. We use SKEWSIZE to evaluate model bias at scale in a variety of domains. We further show how SKEWSIZE can be used with synthetic data to evaluate bias in VLMs.

2 METHOD

Let $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$ be a discriminative model with parameters θ, where $\mathcal{X}$ is the inputs space (e.g. images) and $\mathcal{Y}$ is the label set. We also assume that input $x \in \mathcal{X}$ with label $y \in \mathcal{Y}$ is drawn from an underlying distribution $p(x|z, y)$, where z is a discrete latent variable $z \in \mathcal{Z}$ that represents a factor of variation affecting the data generating process. We refer to z as the *bias* variable and assumed to systematically affect how well the model f_θ is able to predict y from x. Our goal is then estimating to what extent the predictions are affected by z. Let $\mathcal{Z} = \{A, B\}$ and $\mathcal{Y}$ be a discrete set. We further consider that $f_\theta(x)$ defines a conditional distribution $q(y|x; \theta)$ for each $x \in \mathcal{X}$. For a fixed value of $y' \in \mathcal{Y}$, distributional bias should account for systematic differences in the outcomes of $f_\theta(x)$ across different subgroups, i.e. when x is sampled from $p(x|y, z = A)$ versus $p(x|y, z = B)$. More formally, in Equation 1, we define distributional bias as a comparison between *induced families of distributions* defined by $f_\theta(x)$ when $x \sim p(x|y = y', z = A)$ versus when $x \sim p(x|y = y', z = B)$:

$$\mathcal{H}(Q_A(y|x; \theta)||Q_B(y|x; \theta)), \tag{1}$$

where $Q_A(y|x; \theta)$ and $Q_B(y|x; \theta)$ denote the family of distributions obtained when the bias variable assumes each of its possible values, i.e. $z = A$ and $z = B$, respectively. $\mathcal{H}(\cdot||\cdot)$ represents an operator that accounts for a notion of similarity between the two distributions. Depending on the nature of Q, $\mathcal{H}$ can assume different forms. Also, notice that $\mathcal{H}$ operator is not limited to binary attributes and can be instantiated by approaches to compare multiple families of distributions.

As we focus on classification tasks, $f_\theta(x)$ parameterizes families of categorical distributions. We can thus formulate the comparison between Q_A and Q_B as *estimating the effect size* of the association between the bias variable z and the observed model predictions $y' \sim q(y|x, z)$. In this framework, the measure of similarity between Q_A and Q_B can be seen as a measure of association between two categorical variables, the independent variable representing the bias attribute z and y', which we propose to be estimated as per the Cramér's V statistic (Cramér, 1946), and is defined as:

$$\nu = \sqrt{\frac{\chi^2}{N \cdot DF}}, \tag{2}$$

where N the sample size DF is the number of degrees of freedom, and χ^2 represents the test statistic from the corresponding Pearson's chi-squared independence test. Cramér's V is bounded between 0 and 1, with 1 indicating a perfect association between both variables. In order to compute the value of χ^2, the counts of predictions must be arranged in a *contingency table* of size $M = |\mathcal{Z}| \cdot |\mathcal{Y}|$. For a given class y', each entry corresponds to the frequency with each predicted class was observed per subgroup in the data. To obtain a scalar summary metric which can be used to compare multiple models, we propose to aggregate the effect size values using the Fisher-Pearson coefficient of skewness, as it captures both how *asymmetric* the distribution of estimated effect size values is and the *direction* of the asymmetry. For estimated effect sizes $\{\nu_1, \nu_2, \ldots, \nu_{|\mathcal{Y}|}\}$ with empirical mean $\bar{\nu}$, the proposed metric SKEWSIZE is computed as:

$$\text{SKEWSIZE} = \frac{\sum_{i=1}^{|\mathcal{Y}|}(\nu_i - \bar{\nu})^3}{\left[\sum_{i=1}^{|\mathcal{Y}|}(\nu_i - \bar{\nu})^2\right]^{3/2}}. \tag{3}$$

SKEWSIZE can also be implemented considering other choices of statistics, as we show in Appendix F. Here we choose Cramér's V as it is more general and applicable to contingency tables larger than 2x2. Finally, SKEWSIZE can be computed based on logits, softmax scores, top-1 or top-k predictions. Here, we focus on the separation formulation based on top-1 predictions in each class, in which case SKEWSIZE is also applicable to scenarios where this is the only information about the

model's output which is available to the user (Achiam et al., 2023). In Appendix I.4 we propose and empirically validate a strategy to control for noise in the predictions as the output space $|\mathcal{Y}|$ grows, as well as a pseudocode for SKEWSIZE (Algorithm 1) and a Python implementation.

	Accuracy-based			DP(↓)			EO (↓)			**Effect size** (↓)		
Removed	Accuracy (↑)	WG (↑)	GAP (↓)	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2
Unbiased	0.998	0.996	0.002	0.004	0.004	0.001	0.001	0.001	0.001	0.006	0.020	0.024
Class 0	0.888	0.666	0.222	0.050	0.315	0.265	0.038	0.483	0.203	0.705	0.012	0.017
Class 1	0.891	0.653	0.238	0.303	0.013	0.289	0.448	0.009	0.220	0.012	0.703	0.011
Class 2	0.888	0.664	0.224	0.278	0.057	0.332	0.208	0.040	0.484	0.032	0.009	0.700

Table 1: **Effect size captures bias on DSPRITES.** Effect size is only non-negligible for biased classes and indicates which class is affected by spurious correlations.

3 EXPERIMENTS

We empirically demonstrate the effectiveness of SKEWSIZE to measure biases in a controlled experiment with the dSprites (Matthey et al., 2017) dataset and show how it can be applied to assess bias in foundation VLMs from the BLIP-2 family, comparing models in tasks where predicted classes do not necessarily appear as ground-truth in the evaluation dataset. In Appendix G and H we show how SKEWSIZE can provide a better understanding of a model's performance in classification tasks in the IMAGENET (Deng et al., 2009) and DOMAINNET (Peng et al., 2019) datasets, respectively.

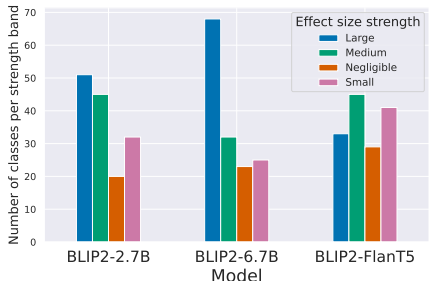
3.1 CONTROLLED SETTING: DSPRITES DATASET

We use DSPRITES dataset considering the task of predicting the object *shape* to evaluate whether the model predictions are biased with respect to the object's *color*. Under a regime of systematic training set manipulation, we induce controlled spurious correlations in the training data by excluding examples of a specific shape and color. This allows us to validate that effect size estimation can be used as a strategy to capture biased predictions and provides information on which class was affected by the introduced spurious correlation. Using the terminology in Section 2, the object color is the *independent variable* (i.e. the variable on which we intervene), and the predicted shape is the *dependent variable* (i.e. the variable we observe). We build three versions of the training data that have different spurious correlations by removing all examples in the GREEN color from one of the classes and train a ResNet18 (He et al., 2016) for 5k steps with each dataset, as well as with the original unbiased training set. Evaluation is carried on test data that have *not* been manipulated.

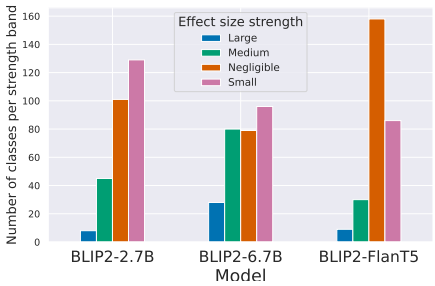
For each ground-truth class, we compute the effect size of the interaction between color and prediction as described in Section 2. We present in Table 1 results in terms accuracy-based metrics (worst-group accuracy (WG) and accuracy gap between subgroups) on an *unbiased* test set, along with Equality of Odds (EO), Demographic Parity (DP) (as defined in Appendix C) and per-class effect size, our approach. We observe that, for all models, effect sizes were strong only for the classes affected by the spurious correlation (i.e. the ones that had green instances removed at training), while remaining negligible for the other classes, confirming that the proposed approach indeed captures model biases and correctly provides per-class granularity. In contrast, EO and DP tend to distribute the effect of this bias across the confused classes, and do not indicate the origins of the confusion. In Appendix E we also show that effect size captures different levels of bias.

3.2 EVALUATING GENDER BIAS IN FOUNDATION VLMS

We now consider the case where the output space is intractable and obtaining data to evaluate the model is challenging. We study the BLIP-2 model family for (binary) gender bias when predicting occupation or practiced sport. Apart from Visogender (Hall et al., 2023b), with only 500 instances, there are no real-world datasets available for evaluating gender biases on VLMs. Therefore, to investigate the utility of SKEWSIZE in the evaluation of VLMs, we generated synthetic data using STABLE DIFFUSION (Rombach et al., 2022) with templates such as: *A* {GENDER} {OCCUPATION}. (see discussion in Appendix B) and query the VLM model with *What is this person's occupation?*. To evaluate the model under conditions that closely resemble their usage "in-the-wild", we *directly*



(a) Gender bias in occupation prediction.



(b) Gender bias in sport modality prediction.

Figure 1: **Gender bias across classes, BLIP2.** Effect size bands: 0-0.1 is negligible; 0.1-0.3, 0.3-0.5, and $>$ 0.5 are small, medium, and large, respectively. Scaling up model size with an unsupervised language model increased total large bias classes, while instruction-tuning decreased it.

use the textual output as predicted class and *do not* constrain the output space of the VLM. More details in Appendix I.5. We investigate models from the BLIP-2 family with different characteristics: BLIP2-2.7B with 3.8B parameters and an unsupervised-trained language model, its larger version BLIP2-6.7B, with 7.8B parameters, and BLIP2-FlanT5), with 4.1B parameters and an instruction-tuned language model. In Table 5 (Appendix I), we report effect size for various occupations in considering predictions by BLIP2-FlanT5. By comparing the accuracy and GAP with effect size for the *Writer*, *Doctor*, and *Biologist* classes, we further validate the main premise of this work. Moreover, we further notice that when GAP is high, the effect size also increases, further showcasing the potential of such a metric to measure disparities between subgroups that also appear as a mismatch between average and worst-case accuracy.

We now compare all three instances of the BLIP2 model family and investigate whether specific characteristics such as increased scale and instruction tuning, amplify or mitigate biases. In Figure 1, we categorize effect size values between 0 and 0.1 as negligible (*negligible* here *does not refer* to the extent that potential harms will affect users) and between 0.1 and 0.3, 0.3 and 0.5, and above 0.5 as small, medium, and large, respectively. For occupation prediction, (Fig. 1(a)), the larger model has more classes which exhibit medium and large effect sizes, suggesting an overall amplification in gender bias. However, using an instruction-tuned language model leads to fewer such classes, suggesting instruction tuning may mitigate bias in this instance. Results for sport modality prediction follow a similar trend (Figure 1(b)). We then compare all models by computing SKEWSIZE . Results repored in Table 2 show that, for both tasks, BLIP2-FlanT5 obtained the highest SKEWSIZE, suggesting that instruction tuning seems to be able to mitigate bias. Moreover, SKEWSIZE values from Table2 further corroborate findings from Fig. 1, as increasing model size amplified biases in predictions.

	Occupation	Sports
BLIP2-2.7B	0.233	1.205
BLIP2-6.7B	-0.045	0.360
BLIP2-FlanT5	0.599	1.255

Table 2: **VLMs SKEWSIZE (higher is better)**. Increasing model size seems to amplify biases, while instruction tuning attenuates it.

4 CONCLUSIONS

We proposed a novel metric, SKEWSIZE, to measure biases in classification models, including when the output space is intractable. Motivated by the observation that certain biases may present in the distribution of prediction errors, we draw on tools from hypothesis testing and propose to measure bias on a per-class basis by estimating the effect size between model prediction and the bias variable. Such an approach allows to obtain a scalar value to compare models as well as detailed information about which are the classes mostly affected by biases. Experiments show that SKEWSIZE captures disparities that accuracy-based metrics do not surface, while not requiring any further information to be computed. Aspects to be investigated in future work include employing SKEWSIZE to evaluate mitigation strategies for neural networks such as (Seth et al., 2023).

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ibrahim Alabdulmohsin, Jessica Schrouff, and Oluwasanmi O Koyejo. A reduction to binary approach for debiasing multiclass datasets. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ogNrYe9CJlH.

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning*. fairmlbook.org, 2019.

Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022.

Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023.

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

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

Harald Cramér. Mathematical methods of statistics, 1946. *Department of Mathematical SU*, 1946.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Cyrus DiCiccio, Sriram Vasudevan, Kinjal Basu, Krishnaram Kenthapadi, and Deepak Agarwal. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1467–1477, 2020.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

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.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. *arXiv preprint arXiv:2301.11100*, 2023a.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *arXiv preprint arXiv:2306.12424*, 2023b.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Fanny Jourdan, Laurent Risser, Jean-Michel Loubes, and Nicholas Asher. Are fairness metric scores enough to assess discrimination biases in machine learning? *arXiv preprint arXiv:2306.05307*, 2023.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Explaining visual biases as words by generating captions. *arXiv preprint arXiv:2301.11104*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 491–507. Springer, 2020.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *European Conference on Computer Vision*, pp. 270–288. Springer, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Christina Lu, Jackie Kay, and Kevin McKee. Subverting machines, fluctuating identities: Re-learning human categorization. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1005–1015, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533161. URL https://doi.org/10.1145/3531146.3533161.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*, 2023.

Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15, 2023.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images. *arXiv preprint arXiv:2305.19164*, 2023.

Preston Putzel and Scott Lee. Blackbox post-processing for multiclass fairness. *arXiv preprint arXiv:2201.04461*, 2022.

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*, pp. 8748–8763. PMLR, 2021.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Julien Rouzot, Julien Ferry, and Marie-José Huguet. Learning optimal fair scoring systems for multi-class classification. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 197–204. IEEE, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6820–6829, 2023.

Sahil Singla, Atoosa Malemir Chegini, Mazda Moayeri, and Soheil Feiz. Data-centric debugging: mitigating model failures via targeted data collection. *arXiv preprint arXiv:2211.09859*, 2022.

Brandon Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. *arXiv preprint arXiv:2305.15407*, 2023.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, pp. 633–645, 2021.

Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 401–416. IEEE, 2017.

Eddie L Ungless, Björn Ross, and Anne Lauscher. Stereotypes and smut: The (mis) representation of non-cisgender identities by text-to-image models. *arXiv preprint arXiv:2305.17072*, 2023.

Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*, 2023.

Angelina Wang and Olga Russakovsky. Directional bias amplification. In *International Conference on Machine Learning*, pp. 10882–10893. PMLR, 2021.

Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 324–335, 2022.

Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*, 2021.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229, 2022.

Robert Wolfe and Aylin Caliskan. American== white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 800–812, 2022.

William Yik, Limnanthes Serafini, Timothy Lindsey, and George D Montañez. Identifying bias in data using two-distribution hypothesis tests. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 831–844, 2022.

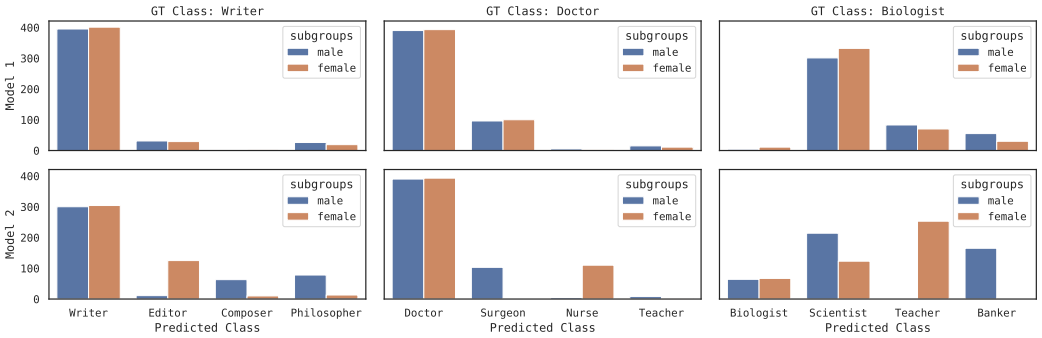
Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35:21682–21697, 2022.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

APPENDIX

A MOTIVATING EXAMPLE



	Writer			Doctor			Biologist			All		
	M1	M2	Δ ↓	M1	M2	Δ ↓	M1	M2	Δ ↓	M1	M2	Δ ↓
Acc	0.88	0.67	-0.21	0.77	0.77	0.00	0.02	0.15	0.13	0.56	0.53	-0.03
WG	0.87	0.66	-0.21	0.77	0.77	0.00	0.01	0.14	0.14	0.55	0.53	-0.02
GAP	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	-0.01	0.01	0.00	0.00
- SKEWSIZE	-0.07	-0.45	**-0.38**	-0.04	-0.45	**-0.40**	-0.12	-0.71	**-0.58**	-0.44	-0.71	**-0.27**

Figure 2: **Standard metrics fail to capture biases within a model.** We plot the prediction counts for two models given three ground-truth classes (*Writer*, *Doctor*, *Biologist*). MODEL 1 (M1) displays similar distributions of errors for both subgroups whereas MODEL 2 (M2) displays 'stereotypical' errors (e.g. mispredicting female *Doctor*s for *Nurse*s). In the table, we report accuracy (Acc), worst group accuracy (WG), GAP and their difference (Δ) between M1 and M2. Only our approach (SKEWSIZE) captures the bias in all settings.

Consider the synthetic setup in Figure 2 which compares two image classification models: MODEL 1 and MODEL 2. These models predict occupation, with different distributions of outputs across two mutually exclusive[2] subgroups (*male* and *female*). Following prior work, we first compute model accuracy in each subgroup (e.g. Chowdhery et al., 2022), worst group accuracy (i.e. minimum accuracy across groups, Sagawa et al., 2019) and GAP (the difference between subgroup accuracy and overall accuracy, Zhang & Ré, 2022) across the following three ground-truth classes:

- WRITER: MODEL 2's accuracy is lower than that of MODEL 1; a bias in MODEL 2's predictions is evident in women being misclassified as *Editor*s and men being misclassified as *Composer*s and *Philosopher*s. Accuracy and Worst group accuracy degrade as expected for the more biased model, whereas GAP does not.
- DOCTOR: Accuracy is the same for MODEL 1 and MODEL 2 but a bias is evident in MODEL 2's predictions, with women being misclassified as *Nurse*s, and men being misclassified as *Surgeon*s. Traditional accuracy-based metrics do not capture this bias.
- BIOLOGIST: Accuracy is higher for MODEL 2 than MODEL 1, but a bias is evident in MODEL 2's predictions, with women being misclassified as *Teacher*s and men as *Scientist*s or *Banker*s. Counterintuitively, the standard metrics improve or stay the same.
- *All*: Aggregating across classes, we can see that the standard metrics either improve in MODEL 2 relative to MODEL 1 or do not change.

In light of this example, we see that regardless of how the performance of a model *in terms of accuracy* varies across subgroups, bias may also arise from systematic errors in incorrect predictions. Importantly, previously proposed metrics do not surface such bias and give the misguided impression that the model's predictions do not exhibit bias.

[2]Assumed to be mutually exclusive for the limited purpose of this illustrative example. We recognize that reality is richer and more nuanced than this binary categorization. See *Impact Statement* section.

B Impact Statement

In this work we propose a metric to estimate how impacted a model is by biases that arise across multiple predictions. We recognize that the binary framing of gender used in the illustrative example and experiments with synthetic data is an oversimplification of an important and complex topic. Our method allows for the interrogation of model bias in terms of discrete, mutually exclusive categories, which may not be ideal for representing multifaceted and intersectional human identities (see Lu et al. (2022) for an exploration of this topic). Finally, the synthetic dataset may inherit stereotypes from its generative model, e.g. misrepresenting non-cisgender people (Ungless et al., 2023). We recommend SkewSize be employed alongside accuracy-based metrics for a more complete picture of a model's performance. We note that SkewSize cannot infer a causal relationship between the bias attribute and model predictions, only their association. We remark that is not within the scope of our work to define which biases are practically relevant, given that this is context-dependent and that a metric should account for all existing biases in a dataset/model so that a comprehensive profile of a model's performance can be taken into consideration at the evaluation.

C Fairness metrics definitions

Previous work on evaluating performance disparity across subgroups has mostly considered metrics such as accuracy (Zhang & Ré, 2022; Alvi et al., 2018; Li et al., 2022), worst group accuracy (Zhang & Ré, 2022; Koh et al., 2021), gap between average and worst group accuracy (referred to as Gap, Zhang & Ré, 2022). These metrics focus on the true positive rate and do not identify biases in the distribution of prediction errors. We compute these metrics throughout the work, for comparison with our approach.

Alternatively, fairness criteria can be formulated as independence desiderata (Barocas et al., 2019), with metrics classified as 'independence' criteria if $f_\theta(x) \perp z$, 'separation' if $f_\theta(x) \perp z|y$ and 'sufficiency' if $y \perp z|f_\theta(x)$. In practice, these criteria are operationalized using different metrics. For the independence criterion, demographic parity (Dwork et al., 2012, DP) is commonly used. These metrics have been recently extended for use in the multiclass setup (e.g. Alabdulmohsin et al., 2022; Pagano et al., 2023; Putzel & Lee, 2022; Rouzot et al., 2022). In this case, metrics are typically computed by binarizing each class (e.g. Alabdulmohsin et al., 2022; Pagano et al., 2023) and aggregating fairness scores across classes using their maximum (i.e. worst case scenario), or average (c.f. Appendix C). Given a full confusion matrix, equality of odds (EO) (Hardt et al., 2016), and potentially DP, would capture differences in the distributions of model errors. However, the detected bias would be surfaced in the scores of the confused classes rather than associated with the class of interest. In our motivating example, EO comparing male and female examples in the Doctor class would be close to 0, but larger for the Surgeon and Nurse classes. In an intractable output space, a full confusion matrix may be unavailable, and EO and DP would be limited in their ability to highlight differences in the distribution of model errors. In this work, we compute EO and DP as per Alabdulmohsin et al. (2022) when a full confusion matrix is available.

We consider Demographic Parity Dwork et al. (2012) as an independence fairness criterion:

$$\text{DP} = \max_{a \in \mathcal{Z}} \mathbb{E}[f_\theta(x) \mid z = A] - \min_{z \in \mathcal{Z}} \mathbb{E}[f_\theta(x) \mid z = a]. \tag{4}$$

While for separation, we refer to equalized odds (EO, Hardt et al., 2016):

$$\text{EO} = \max_{a,k \in \mathcal{Z}x\mathcal{Y}} \mathbb{E}[f_\theta(x) \mid z = a, y = k] - \min_{a,y \in \mathcal{Z}x\mathcal{Y}} \mathbb{E}[f_\theta(x) \mid z = a, y = k]. \tag{5}$$

We focus on the multi-class extension of these metrics by binarizing the task, as suggested in (Alabdulmohsin et al., 2022). The metrics are then aggregated across classes using their maximum value.

D Related Work

Fairness hypothesis testing. Previous work has proposed hypothesis testing approaches to probe for fairness under multiple definitions within datasets (Caliskan et al., 2017; Yik et al., 2022) and

algorithms (Jourdan et al., 2023). Tramer et al. (2017) introduced a permutation test based on Pearson's correlation statistic to test for statistical dependence, under a particular metric, between an algorithm's outputs and protected user groups, while DiCiccio et al. (2020) proposed to test the hypothesis that a model is fair across two groups as per any given metric. Our work differs from both a methodological perspective, e.g. in comparison to Yik et al. (2022) which considers whether the data distribution is significantly different from a reference distribution, as well as applicability, since we propose a metric that can capture biases in a multi-class setting, and which goes beyond binary sensitive attributes (DiCiccio et al., 2020).

Evaluating biases in neural network. Previous work on bias evaluation has prioritized tasks where the information necessary to measure bias can be directly inferred from text (Rae et al., 2021; Wang et al., 2022; Tang et al., 2021; Wang et al., 2021) or by another model (Naik & Nushi, 2023). In contrast, we evaluate bias directly in the model output space, as opposed to relying on predictions of subgroup information. Previous work (Birhane et al., 2023; Luccioni et al., 2023) found that scale appears to amplify stereotyping and bias, as well as reflect biases in the training data (Radford et al., 2021; Wolfe & Caliskan, 2022; Hall et al., 2023a; Prabhu et al., 2023). In the case of VLMs, most prior work focused on leveraging annotated datasets such as MS-COCO (Chen et al., 2015), CelebA (Liu et al., 2015) and FairFace (Karkkainen & Joo, 2021) to measure and mitigate bias (Berg et al., 2022; Chuang et al., 2023; Hall et al., 2023a), while Seth et al. (2023) and Smith et al. (2023) collected a benchmark and obtained synthetic contrast sets, respectively. Prior work (Zhao et al., 2017; Wang & Russakovsky, 2021) has also evaluated bias amplification, but comparing prediction statistics with the original dataset.

Mitigations. Given a known bias in the model, it is possible to mitigate the issue, demonstrating the importance of being able to identify biases to improve the model. This can be done by intervening on the dataset to make it fairer while maintaining performance as done by Singla et al. (2022). Another approach is to intervene on the prompts and de-bias the text embeddings as done by Chuang et al. (2023). Finally, we can intervene at the model level, as done by Friedrich et al. (2023); Berg et al. (2022) and use guidance or an adversarial loss to steer the model towards being more fair. (Zhang et al., 2018), (Alvi et al., 2018) (Kim et al., 2023), (Li et al., 2022)

E Effect Size Captures different bias levels

Following a similar set-up of the dSprites experiment on Section 3.1, we now induce different levels of the same spurious correlation by creating training datasets containing different number of examples from the combination of color and class. We created three datasets by removing instances from Class 1 in the Green color so that only $\{5k, 2k, 0\}$ of such examples are left in the training data. We adopt the same architecture, training, and evaluation from the previous experiment. Results are shown in Table 3, where, for reference, we also report the performance of the unbiased model. As expected, accuracy-based metrics decreased as the number of examples from the removed class, color increased, confirming the models are increasingly affected by the induced spurious correlation. We find that the effect size for the affected class presents a monotonic increasing relationship with bias strength, effect size for the unaffected classes remained negligible, confirming that the effect size captures different levels of bias and correctly indicates the affected class.

	Acc.-based		**Effect size** (↓)		
Bias strength	Acc. (↑)	WG (↑)	Class 0	Class 1	Class 2
Unbiased	0.998	0.996	0.006	0.020	0.024
Mild	0.977	0.936	0.002	**0.253**	0.017
Medium	0.933	0.806	0.013	**0.492**	0.032
Strong	0.891	0.653	0.012	**0.703**	0.011

Table 3: **Inducing varying bias strengths in models trained on dSprites**. The bias strength denotes the number of examples from Class 1 in the green color that were left in the respective versions of the training data. No Bias: full training set, Mild: 5k, Medium: 2k, Strong: 0.

F Computing Effect Size using Other Statistics

In addition to the 3 accuracy based metrics and 2 fairness metrics we already considered in previous results, in this section we further include the Phi coefficient as a measure of effect size when computing SkewSize in the dSprites experiments. The results in Table 4 show that in this case the Phi Coefficient yields similar trends as the Cramer's V. Notice, however, that it is not advisable to use the Phi Coefficient on contingency tables larger than 2x2, which is the reason why we decided to use the more general Cramer's V when computing SkewSize throughout our work.

	Cramer's V			Phi Coefficient		
	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2
Unbiased	0.012	0.011	0.019	0.012	0.011	0.027
Class 0	**0.670**	0.015	0.016	**0.948**	0.015	0.022
Class 1	0.014	**0.683**	0.108	0.014	**0.966**	0.152
Class 2	0.047	0.006	**0.696**	0.067	0.006	**0.985**

Table 4: Computing effect size with Cramér's V vs Phi Coefficient. dSprites dataset.

G Estimating distributional bias in multi-class classification: DomainNet

We now stress-test SkewSize by employing it to evaluate a model in the multi-domain setting, where samples from different distributions are employed training time, and show that our proposed metric can capture systematic biases in predictions. Specifically, we investigate the degree of bias exhibited by the model with respect to the different domains (in this setting, the domain label corresponds to the spurious bias variable).

Setting. We consider the DomainNet benchmark (Peng et al., 2019), which is composed of images from 6 domains sharing the same label space of 345 object classes, and train a ResNet-50 on the train split of all domains jointly. Given the trained model, we then compute predictions for all instances in the test partitions and proceed to compute SkewSize as per Algorithm 1.

Results. The model achieved 59.95% average test accuracy, 37.01% worst group accuracy gap, and 0.509 SkewSize. In order to provide a fine-grained understanding about the differences between each metric, we show in Figure 3 plots accuracy (per class) against effect size ν, along with the respective Equality of Odds (EO) value (shown as each point's corresponding hue). We find a mild Pearson correlation between effect size and accuracy (-0.291, $p \approx 0$) as well as between effect size and EO (0.190, $p = 0.0008$), which indicate the metrics are related but not equivalent as they capture distinct aspects of the bias. No correlation between effect size and Gap was found (0.103, $p = 0.07$), nor between effect size and DP (0.051, $p = 0.377$) further highlighting the importance of including robustness evaluations metrics that take into account error mismatches for a given ground-truth class.

H Estimating distributional bias in multi-class classification: ImageNet

We have thus far demonstrated that SkewSize is capable of accounting for aspects of a model's behaviour that are not captured by accuracy-based bias metrics. We now showcase how SkewSize can be used to provide a more comprehensive evaluation of classifiers by distinguishing models that perform similarly in terms of accuracy, but turn out to display different levels of bias.

Models. We consider models spanning four architectures: ResNet50s (He et al., 2016), Vision Transformers (ViTs) (Dosovitskiy et al., 2020), Inception (Szegedy et al., 2015), and BiT models (Kolesnikov et al., 2020). Architecture and training details are described in Appendix H.1.

Data. We consider a scenario where the background of an image corresponds to the bias variable to evaluate the SkewSize of each model. As no background annotations are available in the original

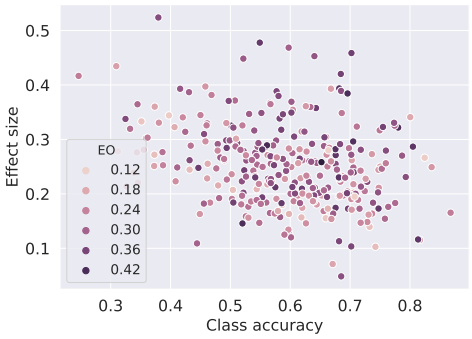


Figure 3: DOMAINNET. Per-class accuracy vs. effect size. Hue indicates EO. Points in the top-right most corner of the plot indicate that even for classes where the model is most accurate systematic differences in predictions across subgroups might exist.

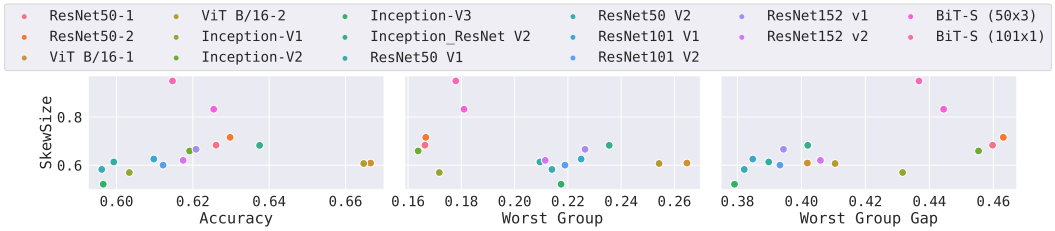


Figure 4: **Comparing models trained on IMAGENET across multiple metrics**. We plot SKEWSIZE versus each accuracy-based metric for a variety of models. The results highlight that no accuracy-based metric presents a clear trend with respect to SKEWSIZE, demonstrating it captures aspects of performance not exposed by these other metrics. Moreover, models with similar performance according to accuracy-based metrics, such as both BiT-S models, can be discriminated by SKEWSIZE .

IMAGENET, we chose 200 classes from the original label set (specifically, those present in TINY-IMAGENET (Le & Yang, 2015)) and generated a synthetic dataset containing images of each of the selected classes across 23 different background types (list obtained from Vendrow et al. (2023)) using STABLE DIFFUSION (Rombach et al., 2022). We generate images using the prompt template *A photo of a* {CLASS} {BACKGROUND} . For instance, for the class SALAMANDER, we used prompts such as *A photo of a* SALAMANDER ON THE ROCKS . We generate 200 images for each background-class pair. Note that these images are used only for evaluation, not training.

Results. In Figure 4 we compare models in terms of accuracy, worst group accuracy, worst group accuracy gap, and SKEWSIZE. The first aspect to observe is that, overall, no clear correlation between these metrics and SKEWSIZE: models with similar accuracy may present considerable disparities in how biased they are as demonstrated by the differences in SKEWSIZE values. Specifically, we highlight that although models such as BIT-S 50X3 and 101X1 present similar performance as per all considered accuracy-based metrics, they can be further discriminated by SKEWSIZE as BIT-S (101X1) achieved higher a value for this metric.

Uncovering spurious correlations with SKEWSIZE. We now examine specific cases of systematic bias uncovered by SKEWSIZE. We identify examples by investigating classes where the model is both accurate and the effect size for the association between background and the model's prediction is high. In Figure 5, we show the top-3 predictions by the VIT B/16-1 for SOCKS in sub-groups corresponding to *A photo of a* SOCK ON THE ROAD and *A photo of a* BLUE SOCK . Both sub-groups/domains present similar measured accuracy, in which case metrics such as worst group accuracy and GAP would be ineffective to capture bias that can be observed in the misclassified

cases. This disparity in the distribution most frequent errors for each subgroup is in fact captured by SKEWSIZE and suggest that the evaluated model may incorrectly associate an ON THE ROAD background with the class RUNNING SHOES, even when the true object of interest is absent.

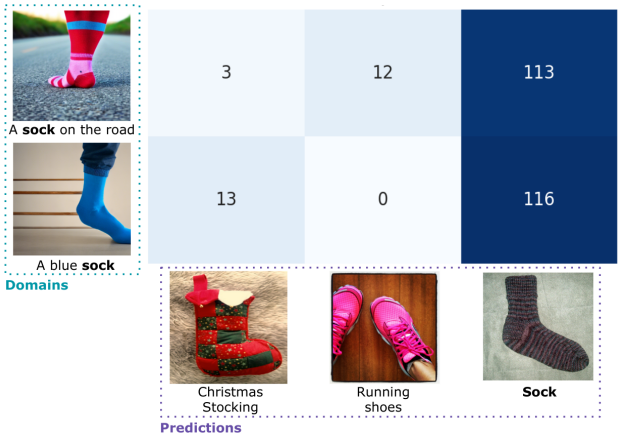


Figure 5: **Bias exposed by SKEWSIZE.** Both domains for the SOCKS class have similar accuracy, but a mismatch in errors indicates the model relies on spurious features of background/color.

H.1 IMAGENET MODELS

We used a variety of models trained on IMAGENET with different sizes, training accuracy, pretraining, etc. Unless otherwise stated, we used publicly available models from TF-HUB[3].

- RESNET50-1/2 (He et al., 2016): A model we trained on IMAGENET from scratch which achieved around 76% accuracy.
- RESNET* (He et al., 2016): RESNET models with no pretraining.
- VIT* (Dosovitskiy et al., 2020): A B/16 variant of the vision transformer model family we trained on IMAGENET from scratch which achieved around 80% accuracy.
- INCEPTION* (Szegedy et al., 2015): Inception models with no pretraining.
- INCEPTION RESNET (Szegedy et al., 2017): A hybrid INCEPTION RESNET model with no pretraining.
- BIT-S* (Kolesnikov et al., 2020): BIT models with no pretraining.

[3]https://tfhub.dev/google/imagenet/

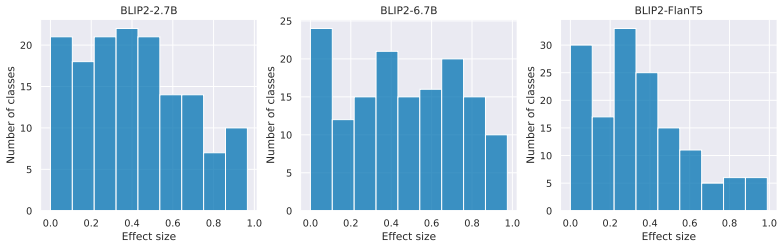
I VLM: DETAILED RESULTS

I.1 ACCURACY-BASED METRICS VS. EFFECT SIZE FOR BLIP2-FLANT5 PREDICTIONS

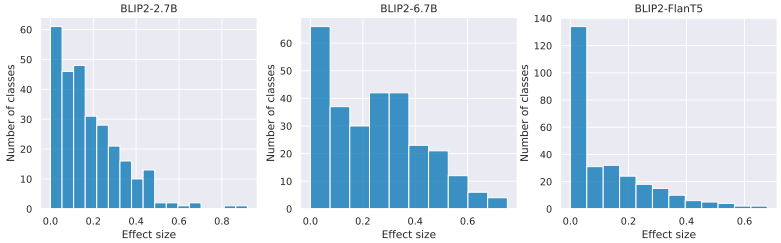
Occupation	Acc. (↑)	GAP (↓)	Effect size (↓)
Writer	0.802	0.006	0.263
Doctor	0.903	0.073	0.291
Biologist	0.151	0.007	0.250
Maid	0.317	0.120	0.556
Model	0.838	0.102	0.368
Nurse	0.517	0.358	0.728
Philosopher	0.349	0.347	0.927
Scientist	0.737	0.065	0.241
Veterinarian	0.791	0.001	0.154

Table 5: **BLIP2-FlanT5.** Even in cases where the GAP is nearly 0, there still is a significant interaction between gender and predicted occupations that accuracy metrics failed to capture.

I.2 EFFECT SIZE DISTRIBUTIONS



(a) Gender bias in occupation prediction.



(b) Gender bias in sport modality prediction.

Figure 6: Distribution of effect size values between gender and predicted occupation/sport modality across BLIP-2 models.

I.3 OPTIONAL POST-PROCESSING

As we do not constrain the model's output, there may be cases where the model predicts synonyms of the ground-truth class, e.g. lawyer and attorney, or the predictions consist of sentences with different structures, e.g. *"The person is a laywer"* and *"A lawyer"*. In light of that, in order to compute accuracy values, we manually post-process the outputs of the model to account for all cases where the output semantically matched the ground-truth answer.

Impact of post-processing. We also investigate in Table 6 whether post-processing model outputs affects the overall experimental findings by comparing the metric trend across models for both raw and post-processed outputs. We find that the same trends can be observed irrespective of the post-processing. Increasing model size while keeping an unsupervised-trained language model amplifies bias as the skewness values decrease when comparing BLIP2-2.7B and BLIP2-6.7B (from 0.233 to -0.045). As expected, SKEWSIZE values computed with raw model outputs tend to be lower, indicating an overall increase in the computed effect size. This is because, without post-processing, the

predicted classes are more fine-grained, resulting in a potential larger mismatch between predictions for each gender value. BLIP2-FlanT5 presented the highest skewness values for all cases, further confirming the findings in Figure 1.

	Raw	Occupation
BLIP2-2.7B	✗	0.233
	✓	-0.005
BLIP2-6.7B	✗	-0.045
	✓	-0.130
BLIP2-FlanT5	✗	0.599
	✓	0.124

Table 6: **SKEWSIZE for raw versus post-processed model outputs**. Higher skewness values correspond to models having less gender bias. We observe that post-processing the models outputs changes the skewness value but *does not* change the overall trend.

I.4 CONTROLLING THE EFFECT OF NOISE IN THE PREDICTIONS

As the size of output space $\|\mathcal{Y}\|$ grows, we propose the following strategy to control for the sensitivity of SKEWSIZE to noise in the predictions: as per the rule-of-thumb to satisfy the assumption of the Chi-square test, we can remove columns from the contingency with respective expected value lower than 5. As we are looking for systematic patterns in the errors of the model, using such a filtering strategy reduces sensitivity to randomness while maintaining sensitivity to the systematic patterns. We can also vary this value in order to decide to which degree some randomness in the predictions should be taken into account.

To illustrate how the choice of the minimum expected value to be accounted for would affect results, we repeated the evaluation reported in Section 3 for the occupation prediction task with varying thresholds so that we can evaluate whether the comparison between models would change. As demonstrated by the results in Table 7, the choice of threshold does not affect the resulting comparison between models.

	MEV=6	MEV=5	MEV=4	MEV=3	MEV=2
BLIP-2.7	0.235	0.233	0.225	0.199	0.19
BLIP-6.7	-0.031	-0.045	-0.056	-0.072	-0.102
BLIP-FlanT5	0.625	0.599	0.578	0.544	0.507

Table 7: Varying the minimum expected value (MEV) for evaluating the BLIP2 model family in the occupation prediction task.

I.5 DATA GENERATION

We consider 148 and 273 classes for the tasks of occupation and sport modality prediction, respectively.

J SKEWSIZE IMPLEMENTATION DETAILS

PSEUDOCODE

Algorithm 1 Computing SKEWSIZE

for $i = 1, 2, \ldots, |\mathcal{Y}|$ **do**
    Get set of model predictions $\hat{Y}^i = \{\hat{y}_k\}$ for all (x_k, y_k, z_k) where $y_k = y_i$
    for $j = 1, 2, \ldots, |\mathcal{Z}|$ **do**
        Build $\hat{Y}^{ij}$, a subset of $\hat{Y}^i$ with instances where $z_k = z_j$
    end for
    Estimate ν_i, the effect size for the i-th class, using Equation 2
end for
Aggregate effect size estimates per class by computing SKEWSIZE as per Equation 3

PYTHON IMPLEMENTATION

```python
# Copyright 2023 The SkewSize Authors. All rights reserved.
# SPDX-License-Identifier: Apache-2.0

import numpy as np
import pandas as pd
import scipy.stats as stats

v_list = []
for label in unique_labels:
    # predictions: predictions for all instances in the class *label*.
    # subgroups: predictions for all instances in the class *label*.
    df = pd.DataFrame({'predictions': predictions,
                       'subgroups': subgroups})
    crosstab = pd.crosstab(df.subgroups, df.predictions)

    chi2 = stats.chi2_contingency(crosstab)[0]
    dof = min(crosstab.shape)-1
    n = crosstab.sum().sum()
    v = np.sqrt(chi2/(n*dof))
    v_list.append(v)

v_values = np.asarray(v_list)
# When a model predicts correctly all examples
# in a given class across all subgroups
# dof=0 and the corresponding v is NaN.
# We remove NaNs before computing skewsize.
v_values = v_values[~np.isnan(v_values)]
skewsize = stats.skew(v_values)
```