Evaluating Model Bias Requires Characterizing its Mistakes

Isabela Albuquerque [1] **Jessica Schrouff** [1] **David Warde-Farley** [1] **Taylan Cemgil** [1] **Sven Gowal** [1] **Olivia Wiles** [1]

Abstract

The ability to properly 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 worst-group accuracy or 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. It can be used in multi-class settings or generalised to the open vocabulary setting of generative 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 trained on synthetic data, vision models trained on IMAGENET, and large scale vision-and-language 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; Solaiman et al., 2023). To evaluate unintended biases in model outputs, existing metrics divide the population (or test set) into subgroups (based on demographic characteristics, how well-represented in the dataset each group is, or another characteristic of significance) and aggregate the e.g. correct and incorrect outputs across those subgroups as in (Sagawa et al., 2019). However, existing metrics 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.

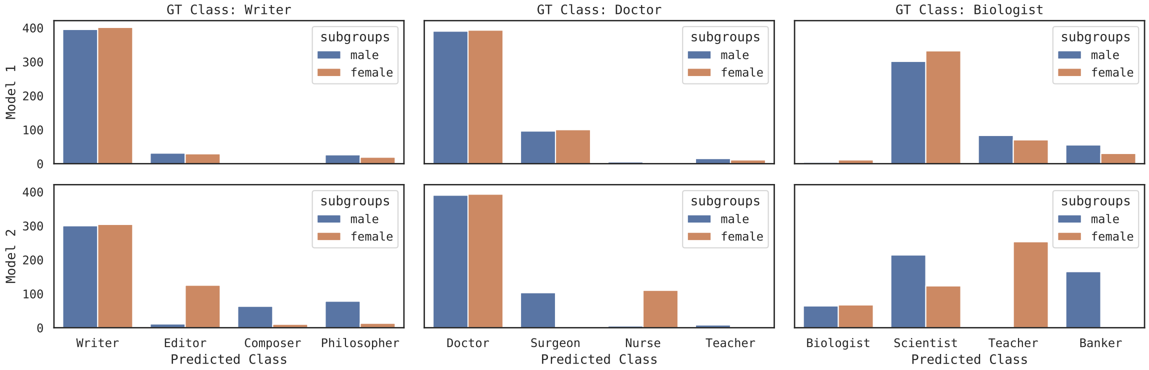
Motivating example. Consider the synthetic setup in Figure 1 which compares two image classification models: MODEL 1 and MODEL 2. These models predict occupation, with different distributions of outputs across two mutually exclusive[1] subgroups (*male* and *female*). Following prior work, we first compute 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.

[1] Google DeepMind. Correspondence to: Isabela Albuquerque <isabelaa@google.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] 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.



	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 1: **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.

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. 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 (e.g. gender) 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: for instance, given images of doctors, we can estimate the effect size corresponding to the association between gender and predicted occupation. As shown in Section 3.1, 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 arrive at a scalar metric which can be used to compare different models. We validate the metric and investigate its utility in three settings: *(1) Controlled data generation:* First, we investigate the utility of the metric, relative to existing ones, by validating it on a controlled setting where spurious correlations affecting single classes are induced, and demonstrate it not only reliably captures bias but also correctly exposes the affected classes. *(2) Multi-class classification:* Next, we demonstrate how our metric can be used to identify previously unidentified cases of systematic bias arising from model mispredictions using DOMAINNET and IMAGENET (Deng et al., 2009; Peng et al., 2019). *(3) Open-ended prediction:* Finally, we analyse large scale vision-and-language models (VLMs) (BLIP-2, (Li et al., 2023)) that have an intractable[2] output space using our metric in two separate settings: gender vs. occupation and gender vs. practiced sport. In settings (2) and (3), we find that no datasets exist that would allow for computing bias metrics among subgroups with statistical significance. As a result, we create synthetic datasets in order to run our evaluation at scale. Our main contributions are summarized as follows:

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 and identify biases arising in the models' errors. We further show how SKEWSIZE can be used with synthetic data to evaluate bias in VLMs.

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

2. Method

2.1. Background

Notation. We consider a discriminative model $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$ with parameters θ, where $\mathcal{X}$ is the set of inputs (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 that affects the data generating process. In the context of this work, z is referred to 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.

Metrics for output disparity across subgroups. 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 A). 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.

2.2. Estimating distributional bias for categorical distributions

Let $\mathcal{Z} = \{A, B\}$ and $\mathcal{Y}$ be a discrete set. We further consider that the parametric model $f_\theta(x)$ 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. $\mathcal{H}(\cdot||\cdot)$ is 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 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*, i.e. the practical relevance of the association between the bias variable z and predictions $y' \sim q(y|x, z)$. In this framework, the similarity between Q_A and Q_B can be seen as a measure of association between two *categorical* variables, z and y'. Given that, effect size can then be estimated via the Cramér's V statistic (Cramér, 1946), 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, i.e. the predictions are extremely sensitive to changes in the value of the bias variable. In order to compute the value of χ^2, the counts of predictions must be arranged in a *contingency table* of size $M = |\mathcal{Z}| \cdot |\mathcal{Y}|$. Contingency tables show the frequency distribution of the variables under consideration, therefore, for a given class y', each entry should correspond to the frequency with each predicted class was observed per subgroup in the data. The χ^2 statistic for such observations accounts for the discrepancy between observed and expected frequencies in the table and is defined as:

$$\chi^2 = \sum_{k=1}^{M} \frac{(o_k - e_k)^2}{e_k}, \tag{3}$$

where o_k refers to the observed value of the k-th entry in the contingency table, and e_k refers to the expected value of this table entry under the assumption of independence between the bias variable and the prediction. In Appendix D.3 we propose and empirically validate an strategy to control for noise in the predictions when computing the quantity in Eq. 3 as the size of output space $|\mathcal{Y}|$ increases.

2.3. Aggregating the Effect Size

The effect size based approach to measure distributional bias evaluates model predictions on a per-class basis. In order to obtain a single, scalar summary metric which can be used to compare multiple models, we must consider how to aggregate the estimated effect sizes for all classes. The ideal metric should be able to simultaneously satisfy the following two conditions: (i) indicate an overall weaker bias when the distribution of effect size values per class is centered around zero with infrequent higher values (as classes for which the model is strongly affected by bias are rare), (ii) distinguish models weakly exhibiting bias from models where, for a considerable fraction of classes, the predictions exhibit high degrees of association with the bias variable, (i.e., the distribution of effect size values is long-tailed and skewed towards the left).

Given the aforementioned desiderata, we propose to aggregate the effect size values per class using the Fisher-Pearson coefficient of skewness, as it captures both how *asymmetric* the distribution of estimated effect size values is as well as 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:

$$\textsc{SkewSize} = \frac{\sum_{i=1}^{|\mathcal{Y}|}(\nu_i - \bar{\nu})^3}{\left[\sum_{i=1}^{|\mathcal{Y}|}(\nu_i - \bar{\nu})^2\right]^{3/2}}. \tag{4}$$

In the Appendix we provide pseudocode for SkewSize (Alg. 1), a Python implementation, and a discussion on modulating the impact low count predictions might have.

Flexibility of SkewSize. As we are interested in surfacing biases in the distribution of model errors, our formulation can be related to a 'separation' fairness criterion. However, this does not preclude the implementation of other criteria. For instance, sufficiency could be implemented by conditioning on the predicted class and using the ground-truth class as the independent variable in the χ^2 test. Similarly, we can implement DP by using the model outputs as the independent variable.

SkewSize can also be implemented considering other choices of statistics, as we show in Appendix B. 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).

3. Experiments

We first empirically demonstrate the effectiveness of SkewSize to measure biases in a controlled experiment with the dSprites (Matthey et al., 2017) dataset. We then proceed to show the usefulness of SkewSize and how it is complementary to accuracy-based metrics in ubiquitous tasks such as classification with the ImageNet (Deng et al., 2009) dataset and the DomainNet benchmark (in Appendix C). We finalize the experimental validation of SkewSize with its application to assess VLMs from the Blip-2 family and demonstrate it can be used to uncover biases and compare models even in cases where predicted classes do not necessarily appear as ground-truth in the evaluation dataset.

3.1. Controlled setting: dSprites dataset

We begin with a setting where controlled levels and types of bias can induced. We use the dSprites dataset, which contains images of objects represented by different shapes, colors and at different positions. We leverage knowledge about the data generating process to introduce spurious correlations in the training data by excluding examples of a specific shape and color. As in this case the effect of relying on such spurious correlations is reflected by accuracy-based metrics, 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.

Setting. We consider the task of predicting the object *shape* and evaluate whether the model predictions are biased with respect to object *color* under a regime of systematic training set manipulation. 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).

Effect Size Captures different types of bias. Given the dSprites' label set has three classes and with color attributes in {`Red`, `Green`, `Blue`}, we build versions of the training data that have different spurious correlations between color and label by removing all examples in the Green color from one of the classes. We then train a ResNet18 (He et al., 2016) for 5k steps with each dataset and evaluate on held out data that has *not* been manipulated to remove any examples. For each ground-truth class, we compute the effect size of the interaction between color and prediction as described in Section 2. For completeness, we

	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 due to introduced spurious correlations on DSPRITES.** Metrics computed on the predictions by ResNet18s trained on 4 versions of DSPRITES show that Effect size is only non-negligible for the corresponding biased classes and indicates which is the class affected by spurious correlations.

verify that each model reaches nearly 100% accuracy on a test set biased in a similar way. These experiments highlight the existence of bias in the model's predictions: accuracy should be nearly 100% for all classes but the one that had examples removed, in which case errors mostly occur when the object is GREEN.

We present in Table 1 results for the three models and one with the same architecture but trained with unbiased data. We report accuracy-based metrics on an *unbiased* test set, along with EO, DP, and per-class effect size (our approach). As intended, models trained with biased data had lower accuracy in comparison with the model trained with unbiased data, suggesting they indeed rely on the introduced spurious correlation. 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 readily indicate the origins of the confusion.

Effect Size Captures different bias levels. Following a similar set-up to previous experiment, 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 2, 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 2: **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.

3.2. 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 E.

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 IMAGENET, we chose 200 classes from the original label set (specifically, those present in TINYIMAGENET (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

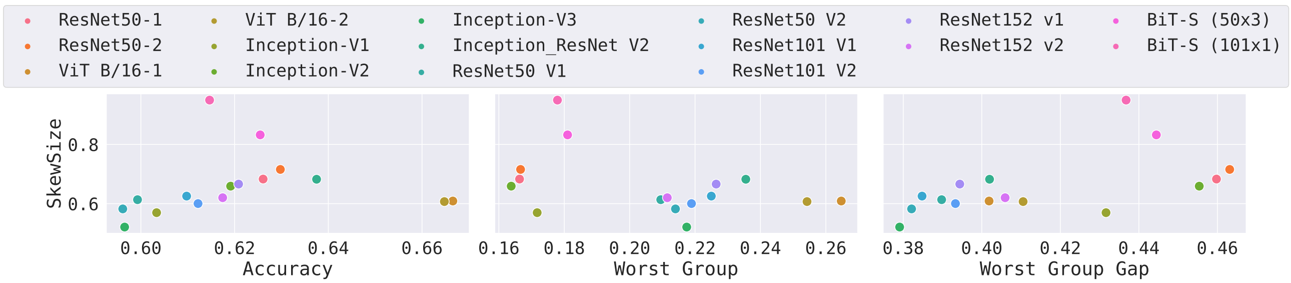


Figure 2: **Comparing models trained on IMAGENET across multiple metrics**. We plot SKEWSIZE versus each accuracy-based metric for a variety of models. The results highlight that no accuracy-based metric presents a clear trend with respect to SKEWSIZE, demonstrating it captures aspects of performance 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 .

that these images are used only for evaluation, not training.

Results. In Figure 2 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 3, we show the top-3 predictions by the VIT B/16-1 for SOCKS in subgroups 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.

3.3. Comparing VLMs for multi-class classification across model size

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.

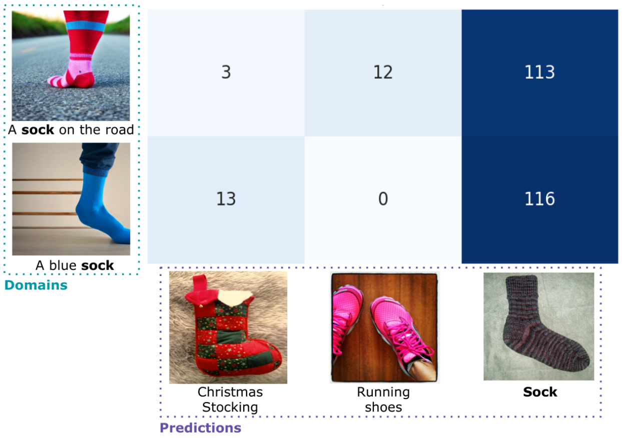


Figure 3: **Bias exposed by SKEWSIZE.** Both domains for the SOCKS class have similar accuracy, but a mismatch in errors indicates the model relies on spurious features of background/color.

Data. 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 gather synthetic data with templates constructed as follows. For the occupation task, we use templates of the form *A* {GENDER} {OCCUPATION}. and query the VLM model with *What is this person's occupation?* . In order to evaluate the model under conditions that closely resemble their usage "in-the-wild", we *directly* use the textual output as predicted class and *do not* constrain the output space of the VLM in order to obtain predictions within the label set of the generated dataset. More details in Appendix D.4.

Models. We again leverage STABLE DIFFUSION (Rombach et al., 2022) to generate data in order to investigate models from the BLIP-2 family with different characteristics (size, instruction tuning). Specifically, we consider

Occupation	Accuracy (↑)	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 3: **VLM evaluation.** Even in cases where the accuracy gap is nearly 0, results show there still is a significant interaction between gender and predicted occupations (e.g. the *Writer* class), indicating the existence of bias that accuracy-based metrics failed to capture.

Blip-2 ViT-g OPT$_{2.7B}$ (BLIP2-2.7B) with 3.8B parameters and an unsupervised-trained language model, Blip-2 ViT-g OPT$_{6.7B}$ (BLIP2-6.7B), its larger version with 7.8B parameters, and Blip-2 ViT-g FlanT5$_{XL}$ (BLIP2-FlanT5), with 4.1B parameters and an instruction-tuned language model.

Results. We report effect size for various occupations in Table 3 considering predictions by BLIP2-FlanT5. By comparing the accuracy and Gap with effect size for the three classes reported in Table 1, namely *Writer*, *Doctor*, and *Biologist*, we further validate the main premise of this work. Results for the remaining classes provide evidence 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.

3.3.1. SkewSize across varying models

We extend the scope of our evaluation and consider all three instances of the Blip2 model family so that we can investigate whether models are biased in different levels, as well as whether specific characteristics such as increased scale and instruction tuning, amplify or mitigate it.

Effect size strength. In Figure 4, we categorize effect size values between 0 and 0.1 as negligible[3] 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. 6(a)), larger models have 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 classes with large effect in comparison to Blip-2.7B and 6.7B, suggesting instruction tuning may mitigate bias in this instance. Results for sport modality prediction follow a similar trend (Figure 4(b)). The number of classes with negligible effect size decreases when increasing model size, while Blip2-FlanT5 exhibits less bias.

SkewSize. As per Section 2, a fine-grained analysis of the results in Figure 4 reveals that the empirical distribution of effect size values (Figures 6(a) and 6(b)) across all classes is heavy-tailed and left-skewed. We report in Table 4 the skewness coefficients for all models in both tasks. We find that larger models seem to exhibit more bias but instruction fine-tuning seems to mitigate the bias.

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

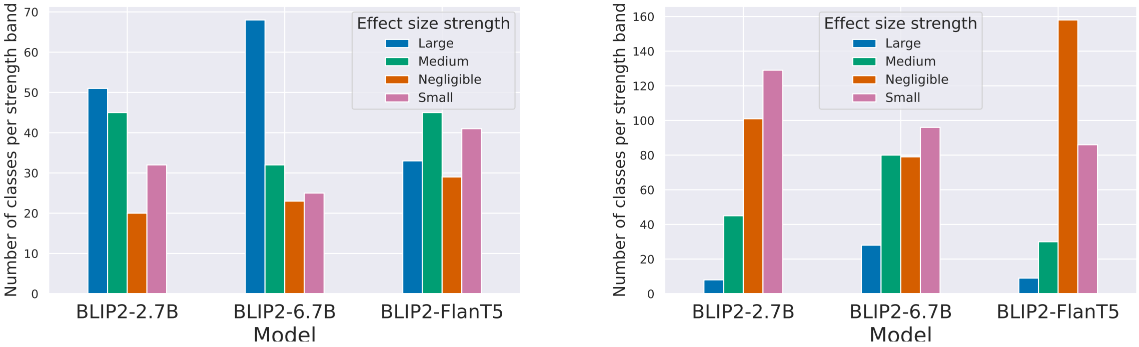
Table 4: **VLMs SkewSize**. Measuring gender bias on occupation and sports modality prediction. Higher SkewSize values are better. Increasing model size seems to amplify biases, while instruction tuning attenuates it.

4. 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., 2022a; 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;

[3] The use of the word *negligible* here *does not refer* to the extent that potential harms will affect users.



(a) Gender bias in occupation prediction.

(b) Gender bias in sport modality prediction.

Figure 4: **Comparing effect size across classes - BLIP2.** Splitting effect size values in bands: 0-0.1 is a negligible effect, while 0.1-0.3, 0.3-0.5, and above 0.5 correspond to small, medium, and large, respectively. Scaling up model size with an unsupervised language model increased the amount of large effect size classes, whereas instruction-tuning decreased it.

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.

5. Discussion

In this work, we propose 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 contingency table hypothesis testing and propose to measure bias on a per-class basis by estimating the effect size of the interaction 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. SKEWSIZE does not require any information besides model's outputs to be computed, therefore not introducing any further requirement in comparison to accuracy-based or fairness metrics.

Experiments across 4 datasets including synthetic examples in cases where no real dataset is available for evaluation, show that SKEWSIZE captures disparities that accuracy-based metrics do not surface. When the full confusion matrix is available, we also highlight that SKEWSIZE complements standard metrics like demographic parity and equalized odds by identifying classes that are affected by spurious correlations. Our results also show how SKEWSIZE can be used in practice: the per-class bias profile yielded by our approach sheds light on spurious correlations for classes presenting both higher accuracy and high effect size, while the aggregated SKEWSIZE is useful to compare models, highlighting, for example, whether increased scale amplifies biases on VLMs, and evaluate the impact of techniques as instruction tuning. Our approach can be declined in a variety of previously established metrics, whether based on accuracy, or on fairness criteria (Barocas et al., 2019). In addition, it caters for binary and discrete bias attributes, and could be extended to continuous attributes by using a different statistic (e.g. Tramer et al., 2017; Brown et al., 2023). Aspects to be investigated in future work include employing SKEWSIZE to evaluate previously proposed mitigation strategies for bias in neural networks such as (Seth et al., 2023), and developing strategies for debiasing.

Practical Considerations. We recommend SKEWSIZE be employed alongside accuracy-based metrics for a more complete picture of a model's performance. Depending on the goal of the evaluation, either per-class measurements can be used or the overall SKEWSIZE. In case multiple models are under comparison and the problem involves multiple classes, we recommend using SKEWSIZEso that a ranking for the models can be obtained using this overall notion of bias. In cases a specific class is investigated or a fine-grained profile of model bias is required, per-class effect size can be employed. We further note that SKEWSIZE cannot infer a causal relationship between the bias attribute and model predictions, only their association. As with mathematical fairness criteria, our metric does not relate bias to potential harms (Weidinger et al., 2022); further work is needed to understand the impact of this distributional bias. Finally, 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.

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).

Acknowledgements

We thank Ira Ktena, Lisa Anne Hendricks, Canfer Akbulut, Shakir Mohamed, Simon Osindero, and Florian Stimberg for their feedback throughout the project. We are grateful to Lisa Anne Hendricks and Simon Osindero for their feedback on the manuscript.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Alvi, M., Zisserman, A., and Nellåker, C. 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.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and machine learning*. fairmlbook.org, 2019.

Berg, H., Hall, S. M., Bhalgat, Y., Yang, W., Kirk, H. R., Shtedritski, A., and Bain, M. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022.

Birhane, A., Prabhu, V., Han, S., and Boddeti, V. N. On hate scaling laws for data-swamps. *arXiv preprint arXiv:2306.13141*, 2023.

Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., and Schrouff, J. Detecting shortcut learning for fair medical AI using shortcut testing. *Nat. Commun.*, 14(1):4314, July 2023.

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

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chuang, C.-Y., Jampani, V., Li, Y., Torralba, A., and Jegelka, S. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

DiCiccio, C., Vasudevan, S., Basu, K., Kenthapadi, K., and Agarwal, D. Evaluating fairness using permutation tests. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1467–1477, 2020.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. 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.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. 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.

Friedrich, F., Schramowski, P., Brack, M., Struppek, L., Hintersdorf, D., Luccioni, S., and Kersting, K. Fair diffusion:

Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

Hall, M., Gustafson, L., Adcock, A., Misra, I., and Ross, C. Vision-language models performing zero-shot tasks exhibit gender-based disparities. *arXiv preprint arXiv:2301.11100*, 2023a.

Hall, S. M., Abrantes, F. G., Zhu, H., Sodunke, G., Shtedritski, A., and Kirk, H. R. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *arXiv preprint arXiv:2306.12424*, 2023b.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (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.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jain, S., Lawrence, H., Moitra, A., and Madry, A. Distilling model failures as directions in latent space. *arXiv preprint arXiv:2206.14754*, 2022.

Jourdan, F., Risser, L., Loubes, J.-M., and Asher, N. Are fairness metric scores enough to assess discrimination biases in machine learning? *arXiv preprint arXiv:2306.05307*, 2023.

Karkkainen, K. and Joo, J. 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.

Kim, Y., Mo, S., Kim, M., Lee, K., Lee, J., and Shin, J. Explaining visual biases as words by generating captions. *arXiv preprint arXiv:2301.11104*, 2023.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. 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.

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

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

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

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

Lu, C., Kay, J., and McKee, K. 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.

Luccioni, A. S., Akiki, C., Mitchell, M., and Jernite, Y. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

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

Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., Araujo, M. M., Santos, L. L., Cruz, M. A., Oliveira, E. L., 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.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Porgali, B., Albiero, V., Ryda, J., Ferrer, C. C., and Hazirbas, C. The casual conversations v2 dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10–17, 2023.

Prabhu, V., Yenamandra, S., Chattopadhyay, P., and Hoffman, J. Lance: Stress-testing visual models by generating language-guided counterfactual images. *arXiv preprint arXiv:2305.19164*, 2023.

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Rojas, W. A. G., Diamos, S., Kini, K. R., Kanter, D., Reddi, V. J., and Coleman, C. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Rouzot, J., Ferry, J., and Huguet, M.-J. 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.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Seth, A., Hemani, M., and Agarwal, C. 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.

Singla, S., Chegini, A. M., Moayeri, M., and Feiz, S. Data-centric debugging: mitigating model failures via targeted data collection. *arXiv preprint arXiv:2211.09859*, 2022.

Smith, B., Farinha, M., Hall, S. M., Kirk, H. R., Shtedritski, A., and Bain, M. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. *arXiv preprint arXiv:2305.15407*, 2023.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Daumé III, H., Dodge, J., Evans, E., Hooker, S., et al. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*, 2023.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Tang, R., Du, M., Li, Y., Liu, Z., Zou, N., and Hu, X. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, pp. 633–645, 2021.

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., Juels, A., and Lin, H. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 401–416. IEEE, 2017.

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

Vendrow, J., Jain, S., Engstrom, L., and Madry, A. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*, 2023.

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

Wang, A., Barocas, S., Laird, K., and Wallach, H. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 324–335, 2022a.

Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., and Russakovsky, O. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7):1790–1810, 2022b.

Wang, J., Liu, Y., and Wang, X. E. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*, 2021.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., 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.

Williams, A., Nangia, N., and Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

Wolfe, R. and Caliskan, A. 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.

Yik, W., Serafini, L., Lindsey, T., and Montañez, G. D. 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.

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

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

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

Appendix

A. Fairness metrics definitions

We consider Demographic Parity (DP, Dwork et al., 2012) and equalized odds (EO, Hardt et al., 2016) as per their multi-class extension described in Alabdulmohsin et al. (2022). Each class is binarized (one versus all) before the computation of the metric, and the results are aggregated across classes using their maximum value. The metrics refer to a "maximum gap" between subgroups. For DP, this would mean computing the proportion of positive predictions for a class in each subgroup, comparing the highest with the lowest values across groups:

$$\mathrm{DP}(f_\theta) = \max_{a \in \mathcal{Z}} \mathbb{E}[f_\theta(x) \mid z = a] - \min_{a \in \mathcal{Z}} \mathbb{E}[f_\theta(x) \mid z = a], \tag{5}$$

$$\mathrm{EO}(f_\theta) = \max_{a,k \in \mathcal{Z} \times \mathcal{Y}} \mathbb{E}[f_\theta(x) \mid z = a, y = k] - \min_{a,k \in \mathcal{Z} \times \mathcal{Y}} \mathbb{E}[f_\theta(x) \mid z = a, y = k]. \tag{6}$$

B. 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 5 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 5: Computing effect size with Cramér's V vs Phi Coefficient. DSPRITES dataset.

C. 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 5 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.

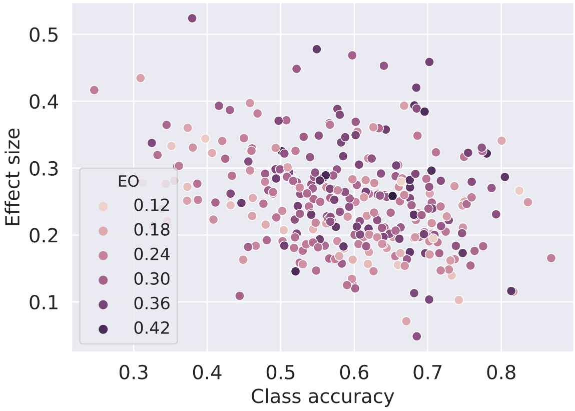


Figure 5: DOMAINNET. Per-class accuracy vs. effect size. Hue indicates EO. Points in the top-right most corner of the plot indicate that even for classes where the model is most accurate systematic differences in predictions across subgroups might exist.

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

Table 6: **SKEWSIZEfor 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.

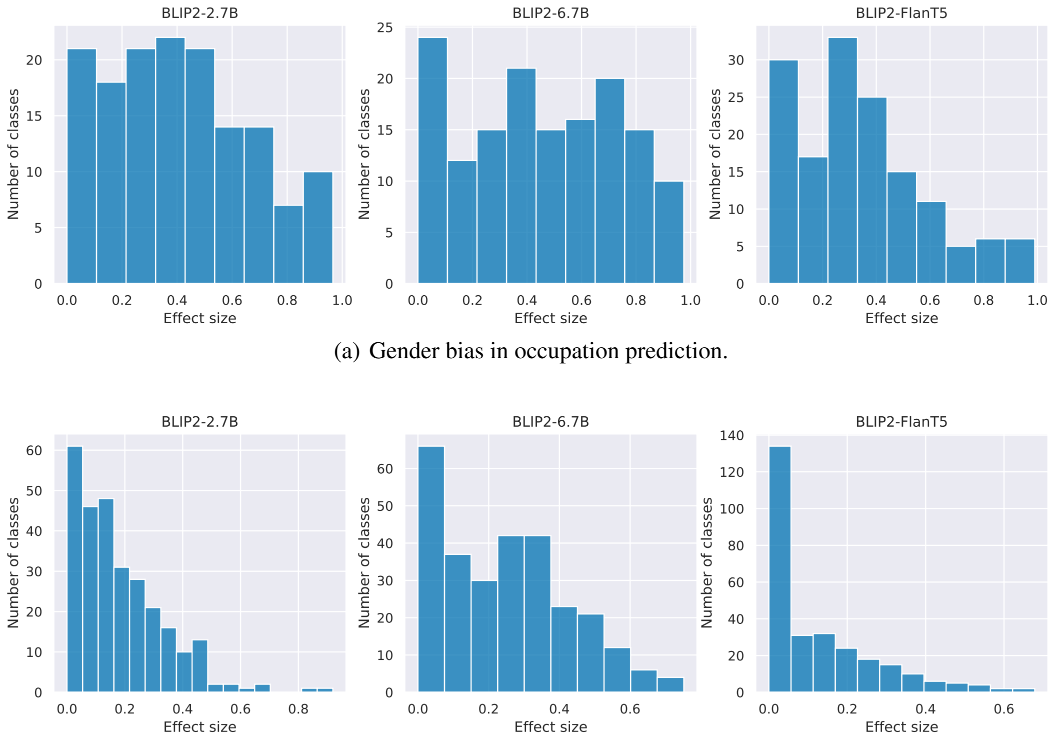
D. VLM: Detailed results

D.1. 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 4.

D.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.

D.3. 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 3.3.1 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.

D.4. Data Generation

We consider 148 and 273 classes for the tasks of occupation and sport modality prediction, respectively. The complete list of occupations and sport modalities used in the VLMs experiments can be found in the Supplementary Material repository.

E. 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[4].

- 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.

F. Further Related Work

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)

Datasets for evaluating bias. To develop methods to evaluate and mitigate bias, datasets such as Waterbirds (Sagawa et al., 2019), CelebA (Liu et al., 2015), and MultiNLI (Williams et al., 2017) have been used; in these datasets the biases are created a-priori (e.g. land birds on land backgrounds vs water birds on watery backgrounds). Revise was introduced by (Wang et al., 2022b) in order to visualise the biases in a dataset (and thereby probable impacts on models trained on such a dataset). However, such a tool requires labels on what exists in the dataset, which may not be possible, so (Jain et al., 2022) demonstrated how biases could be found automatically in multimodal datasets. A targeted dataset looking at model performance at predicting everyday objects conditioned on geographical location is the DollarStreet dataset (Rojas et al., 2022). Other targeted datasets, such as FairFaces and CasualConversationsV2 (Karkkainen & Joo, 2021; Porgali et al., 2023; Hall et al., 2023b) can be used to evaluate a models bias across sensitive attributes. (Karkkainen & Joo, 2021; Porgali et al., 2023) does so by comparing classification performance across these attributes (e.g. age, gender, etc.) whereas (Hall et al., 2023b) is a small dataset of 250 images that evaluates pronoun resolution conditioned on the image. Such datasets are complementary to our approach, as we rely on a dataset with labelled subgroups to compute our metric. When no such dataset exists, we relied on synthetic data.

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

G. SkewSize implementation details

G.1. 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 $\nu_i$, 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 4.
```

G.2. 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)
```