## A. Salience Calculation

Following from Ramaswamy et al. [39], we calculate the relative salience between two attributes, `A` and `B` in the following manner. We sample a dataset from CelebA where half the attributes have both `A` and `B`, and half have neither. We then train a model to perform binary classification on this dataset. Then, we test this model on a test set composed of images equally sampled from the four possible conjunctions of the two attributes. We calculate the AUC on attribute `A` and `B` independently, and take the difference between their performance to indicate which attribute is more salient. We repeat this experiment on a second training dataset skewed for the inverse of `A`, i.e., the two halves are `A` and not-`B`, and not-`A` and `B`.

## B. Correlation Level, Salience, Number of Finetuning Samples on COCO

In Sec. 4.1 we demonstrated how on the CelebA dataset the three factors of correlation level, salience, and number of finetuning images impact whether bias from a pretrained model in the form of a *spurious correlation* will propagate into a finetuned one. Here, we show how these same factors are relevant on the more complex COCO dataset.

In creating our pretrained models, we use the OpenImages dataset [27], which has 600 labels such as `Ladder` and `Carrot` that are annotated by a combination of humans and machines. For gender we follow from prior work [65, 59] and derive these labels based on the presence of gendered labels in the original dataset.[13] We use the subset of this dataset which contains people of only one annotated gender.

Like in CelebA, we use two types of pretrained models: **Gendered** and **Control**. In this setting, our **Gendered** model is pretrained to classify binary gender on OpenImages. Our **Control** model is trained to classify `outdoor parks` or not on OpenImages. These scene labels come from a Places scene classifier [66].

We work with four versions of COCO as our downstream task. Two have varying levels of correlation strength, and two have varying levels of salience. We would expect that on the dataset with more correlation strength compared to less, there would be a bias difference between the two pretrained models; we would not expect as large of a difference on the dataset with less correlation strength. The same hypothesis holds for the two datasets of different salience. To create two versions of COCO that have different levels of correlation strength between the target task (i.e., objects) and sensitive attribute (i.e., gender) we train a logistic regression model to predict gender from a binary vector representing all of the objects present in an image. We then sort all images in the dataset by those most correctly classified by this model to those least correctly classified. We split the images in half to create two datasets, the first we call "Skew_MoreBias" (object presence is highly correlated with gender) and the second set "Skew_LessBiass" (object presence is less correlated with gender). To create two versions of COCO where the salience of the target task and sensitive attribute differ, we blur all of the objects to create "Salience_MoreBias" and blur all of the people to create "Salience_LessBias." As our measure of bias we use the directional bias amplification measure ($\text{BiasAmp}_{\rightarrow}$) from Wang and Russakovsky [60].

In Fig. 7 we show results from finetuning our **Gendered** and **Control** pretrained models on all four variations of the COCO dataset on both 1,000 and 10,000 finetuning samples. When we first compare the results of the two pretrained models on "Skew_MoreBias" and "Skew_LessBias" we see that on the former dataset it makes a difference which pretrained model is used, while for the latter it does not. Somewhat unexpectedly, when we increas the number of finetuning samples from 1,000 to 10,000, the difference between the two pretraining bases increases rather than decreases, as we saw in CelebA. We hypothesize this is because the dataset we have created is so skewed that it benefits the model significantly to continue to learn the spurious correlations, even as the finetuning number has increased.

We see the same results in Fig. 7 for salience where on "Salience_MoreBias" there is a higher difference in directional bias amplification between the two different pretrained bases. However, here we see this gap reduces with additional finetuning samples.

## C. Additional Results from "4.2 Bias from spurious correlations can be corrected for in finetuning"

In Sec. 4.2 we showed results from manipulating the correlation level of the finetuning dataset on both CelebA and COCO. Here, we show additional results on a larger set of downstream tasks for each dataset, as well as different numbers of finetuning samples. In Fig. 8 we show results with 128 and 1024 finetuning samples on the four CelebA attributes which exhibit bias transfer from the pretrained models to the finetuned one. Just like the results we show in the main text, there is a version of each finetuning dataset such that the distribution has a different correlation level than the test dataset, but the performance is retained while fairness improves.

In Fig. 9 we show results with 1000 and 5000 finetuning samples on the four COCO objects which are most represented with both genders. Again, like in the main text, we see that there exists versions of the finetuning dataset that

[13] Schumann et al. [48] has collected a set of more inclusive gender presentaiton labels on this dataset, but we did not use them because the subset these labels exist for was not large enough for our purposes of pretraining.

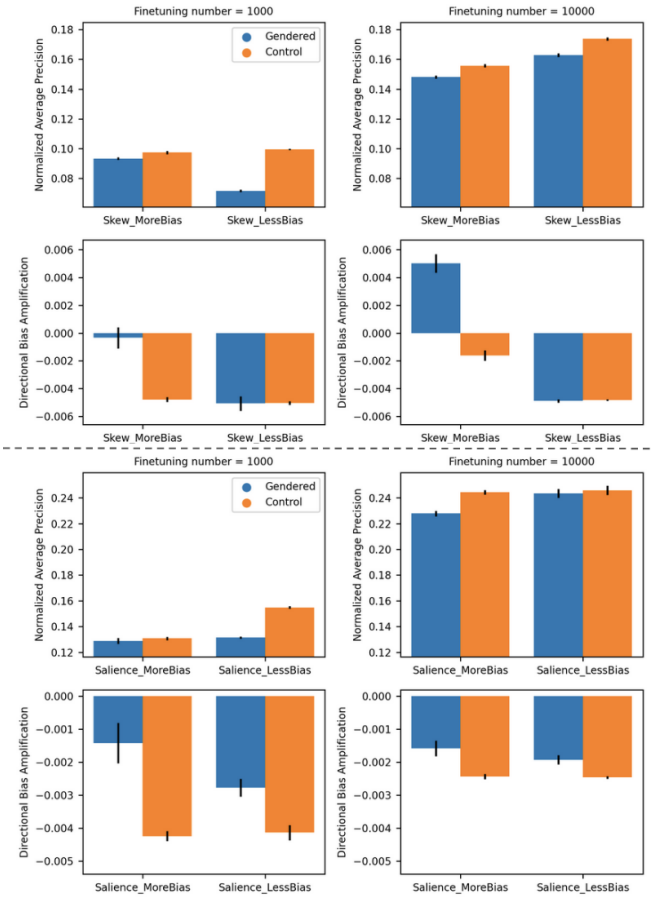


Figure 7. We present results for the performance (normalized average precision) and bias (directional bias amplification) of finetuned models pretrained on two bases: **Gendered** and **Control**. We show results on four different datasets, "Skew_MoreBias" compared to "Skew_LessBiass," and "Salience_MoreBias" compared to "Salience_LessBias."

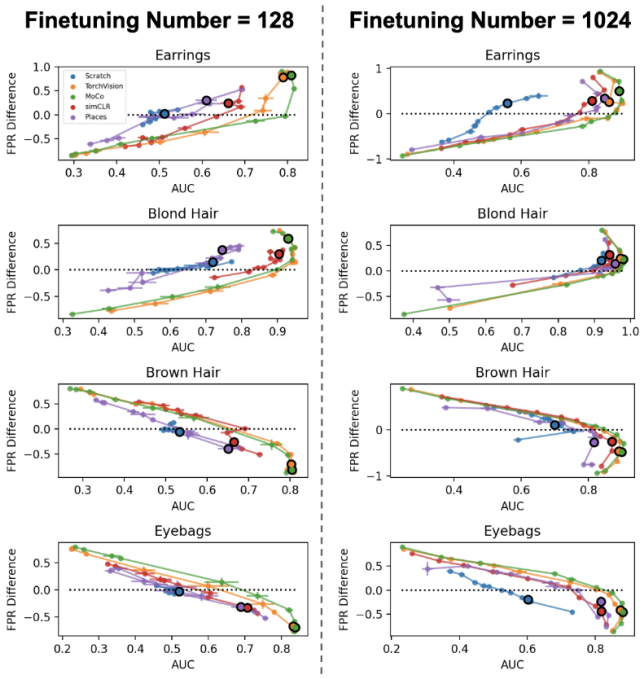


Figure 8. The performance and fairness with 95% confidence intervals of pretrained models finetuned on different versions of four downstream tasks on CelebA: `Earrings` and `Blond Hair` (correlated with women) and `Brown Hair` and `Eyebags` (correlated with men). The bolded point indicates when the finetuning distribution matches the test distribution, and all other points indicate variations on the finetuning dataset. There are versions of the finetuning dataset that allow us to retain performance gains and improve fairness.

allow us to preserve the high performance of a more biased model while decreasing the bias.

## D. Additional Results from "5.1 Finetuned models do worse on subcategories underrepresented in pretrained models"

In Sec. 5.1 we showed on CelebA that finetuned models inherit biases in the form of underrepresentation from pretrained models. Here, we provide further details about our experimental setup, as well as more detailed results disaggregated by correlation level, salience, and finetuning number.

As we had described in our setup, we consider our downstream task `Target` to be composed of two possible subcategories: `T1` and `T2`. We have two possible pretrained models: **Pretrain-T1** that has only been trained to classify `T1`, and **Pretrain-T2** that has only been trained to classify `T2`. Our measure of bias is AUC on `T2` between **Pretrain-T2** and **Pretrain-T1**.

For any instantiation of `T1` and `T2` using CelebA attributes, **Pretrain-T1** and **Pretrain-T2** are trained on their respective attributes on the FairFace dataset. FairFace does not contain attribute labels, so these are labeled by our best classifier which was originally trained on CelebA. While imperfect, we believe this will still provide sufficient training signal for each pretrained model.

We consider three relevant factors which are analogous to those we considered in Sec. 4.1 of correlation level, salience, and finetuning number. For correlation level, we consider the proportion of positive labels that are in subcategory `T1` as compared to `T2`. For salience, we consider the relative salience of `T1` and `T2`. Finetuning number remains the same. In picking attribute pairs to use as `T1` and `T2`, we sample from three discretized types of salience relationships: `T1` is more salient than `T2`, `T1` is equally salient to `T2`, and `T1` is less salient than `T2`. We arbitrarily pick four attribute pairs from each category, and thus look at 12 pairs.

In establishing that finetuned models can inherit biases of underrepresentation from pretrained models, we consider

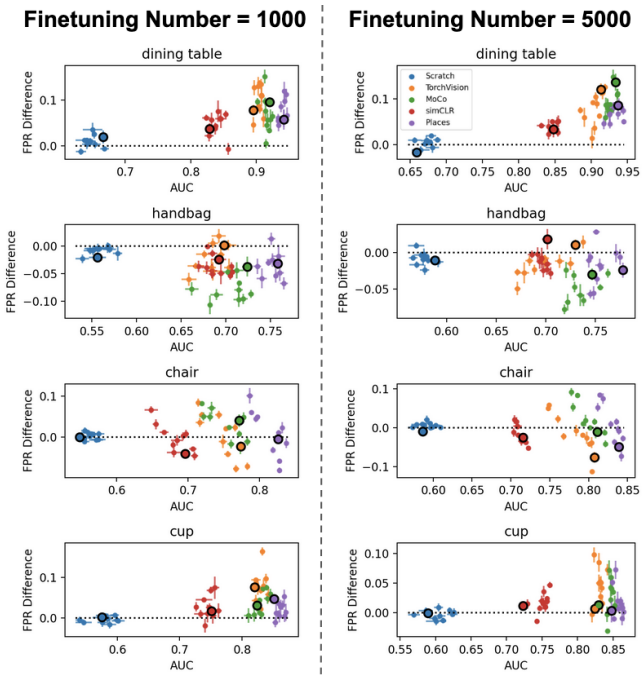


Figure 9. The performance and fairness with 95% confidence intervals of pretrained models finetuned on different versions of four downstream tasks on COCO: `dining table` and `handbag` (correlated with women) and `chair` and `cup` (correlated with men). The bolded point indicates when the finetuning distribution matches the test distribution, and all other points indicate variations on the finetuning dataset. We can see that across both numbers of finetuning there are versions of the finetuning dataset that allow us to retain performance gains and improve fairness.

when the downstream dataset is 50% `T1` and 50% `T2`; the same results on additional proportions are shown in full in Tbl. 1. We do not find clear trends in performance difference on `T2` across the three possible salience relationships, but we do for finetuning number. Across the 12 attribute pairs when we finetune on 128 images, the difference in AUC is $.124 \pm .023$, whereas when we finetune on 1024 images, the difference is $.036 \pm .007$. As expected, increasing numbers of finetuning samples erodes the difference between the different pretrained bases. However, in both cases there is a statistically significant positive difference indicating that a finetuned **Pretrain-T1** is not able to reach the performance on `T2` that a finetuned **Pretrain-T2** is. Even though we do not observe any difference between the different settings of salience, we continue all experiments in the main text across these 12 pairings for generalizability.

Table 1. The top table represents when the finetuning number is 128, and the bottom when it is 1024.

| Correlation Strength / Salience | T1 less than T2 | T1 equal to T2 | T1 more than T2 |
|---|---|---|---|
| 10% | $.17 \pm .06$ | $.06 \pm .03$ | $.13 \pm .09$ |
| 50% | $.15 \pm .07$ | $.07 \pm .02$ | $.15 \pm .09$ |
| 90% | $.10 \pm .06$ | $.07 \pm .03$ | $.07 \pm .09$ |

| Correlation Strength / Salience | T1 less than T2 | T1 equal to T2 | T1 more than T2 |
|---|---|---|---|
| 10% | $.04 \pm .02$ | $.00 \pm .00$ | $.03 \pm .02$ |
| 50% | $.03 \pm 0.02$ | $.03 \pm .01$ | $.05 \pm 0.03$ |
| 90% | $.05 \pm 0.04$ | $.02 \pm .01$ | $.14 \pm 0.05$ |