Appendix

In the supplementary material, we provide additional details on some sections of the main paper.

Sec. A. Additional explanation for our constraint on artifact selection.

Sec. B. Details on how gender labels are automatically derived.

Sec. C. Additional gender artifact model training details.

Sec. D. Additional experiments regarding contextual objects described in Sec. 6 in the main paper.

Sec. E. Additional details on the person and background occlusions described in Sec. 5 in the main paper.

Sec. F. Results from training and evaluating on a gender balanced dataset because the original datasets skewed male.

Sec. G. Additional results on our fairness-through-blindness case study.

A. Constraints on artifact selection

We analyze gender artifacts that are learnable (i.e., result in a learnable difference for a machine learning model) and interpretable (i.e., have an interpretable human corollary). We implement the constraint that the gender artifacts must be interpretable as artifacts perceptible to humans often have the most pressing fairness concerns (e.g., an imperceptible artifact such as a correlation between the n-th pixel in the image and gender may not have as high pressing fairness concerns as if there was a high correlation between outdoor backgrounds and male gender labels). However, we acknowledge there can be infinitely many potential correlations in an image and this criterion is non-exhaustive.

B. Automatically deriving gender labels

Following prior work [76, 83], we use the captions from the Common Objects in Context (COCO) to derive gender labels (Sec. 3). Concretely, we first convert the captions to lowercase. Then, using the following gendered set of words from Zhao *et al.* [82], we query our captions for the presence of these keywords: [“male,” “boy,” “man,” “gentleman,” “boys,” “men,” “males,” “gentlemen”] and [“female”, “girl”, “woman”, “lady”, “girls”, “women”, “females”, “ladies”]. We assign the respective gender label if two of the five captions contain a gendered keyword and discard images for which the captions contain both male and female keywords. We choose to use two of the five captions as these automatically derived captions match the explicit annotations from labelers 85.4% of the time [82]. This suggests label noise for gender labels in the COCO dataset does not significantly affect our results.

	COCO	OpenImages
Full	93.4 ±0.1	81.2 ±0.2
Full NoBg	92.6 ±0.1	-
MaskSegm	79.5 ±0.2	-
MaskRect	70.7 ±0.1	63.3 ±0.4
MaskSegm NoBg	76.0 ±0.1	-
MaskRect NoBg	58.3 ±0.2	59.7 ±0.7

Table 2. **Performance of model trained using five random seeds.** We report the mean AUC (%) and a 95% confidence interval of five gender artifact models trained using random seeds.

C. Additional model training details

Gender artifact model. The model is optimized with stochastic gradient descent (SGD) with a momentum of 0.9 and a batch size of 64. The ResNet-50 has a final fully connected layer mapping the 2048 size hidden layer to a single output value, and we arbitrarily assign “male” as equal to 0 and “female” to 1. We optimize the hyperparameters for each model on the validation set using grid search (learning rate: $\{10^{-2}, 10^{-3}, \dots 10^{-5}\}$; weight decay: $\{10^{-2}, 10^{-3}, \dots 10^{-5}\}$). For the baseline model, input images are resized to 224 x 224 and randomly flipped horizontally during training. Furthermore, we bootstrap until convergence (5,000 resamples) and report a 95% confidence interval on the test set.

Training models on random seeds. We provide the 95% confidence intervals through bootstrapping until convergence (5,000 resamples) for all of results. As an alternative means for providing confidence intervals, for all ResNet-50 models, we train five models on random seeds and provide the 95% confidence intervals as well. As seen in Tbl. 2, 3, and Fig. 6 the intervals are similar to those found via bootstrapping. When removing random contextual objects, we also report the result of training five separate random classifiers as displayed in Fig. 8. In all cases, the gender artifact model continues to perform above random chance.

Pre-training on different datasets. In addition to pre-training on ImageNet as in Sec. 5, we evaluate a ResNet-50 pre-trained on Places-365 [84] and PASS MoCo-v2 [2] weights. We do not train a model from scratch as there is insufficient data. All models perform above random chance (Tbl. 3). In particular, the PASS dataset does not contain any people, indicating that the difference observed is attributable to differences in the input distributions, not the dataset on which the model was pre-trained.

D. Additional contextual object experiments

Additional contextual object results. For COCO, in addition to reporting the results from bootstrapping, we

	ImageNet	Places-365	PASS
Full	93.4 ±0.2	88.4 ±0.2	89.9 ±0.2
Full NoBg	92.7 ±0.2	88.5 ±0.2	89.9 ±0.2
MaskSegm	79.6 ±0.3	74.6 ±0.3	75.8 ±0.3
MaskRect	70.8 ±0.3	69.5 ±0.4	68.6 ±0.4
MaskSegm NoBg	74.8 ±0.3	70.9 ±0.4	74.1 ±0.3
MaskRect NoBg	58.0 ±0.4	58.4 ±0.4	58.8 ±0.4

Table 3. **Pre-training on different datasets.** We report the AUC (%) of a gender artifact model that is pre-trained on one of three datasets: ImageNet [59], Places-365 [84], and PASS [2]. The model is then trained and evaluated on different occlusions (introduced in Sec. 5) on the COCO dataset.

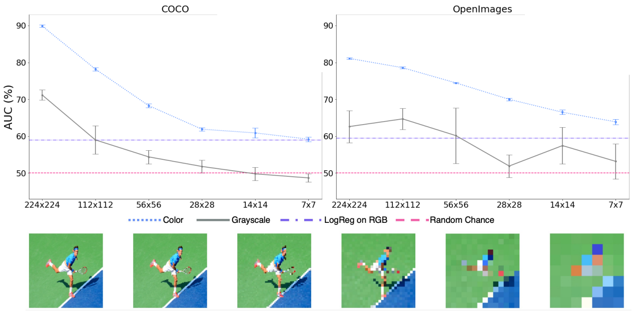

Figure 6. **AUC of models trained with varying resolution and color of input images.** We display another method of calculating confidence intervals for the ResNet-50 model by training five models on random seeds and provide the 95% confidence interval.

Male	Female
Trumpet	Break
Weapon	Tart
Bathtub	Shotgun
Billboard	Dairy Product
Sombrero	Goat
Tiara	Ambulance
Ceiling Fan	Duck
Scoreboard	Banana
Missile	High Heels
Cupboard	Bow and Arrow

Table 4. **Relevant contextual objects.** The ten most relevant objects in descending order, as identified by the weights of the logistic regression classifier trained on OpenImages.

report the results from training five separate random classifiers and report the standard deviation. See Fig. 8 for results.

Next, for OpenImages, in Tbl 4, we report the 10 most relevant objects in descending order, as identified by the weights of the logistic regression classifier trained on OpenImages.

Visualizing gender artifacts through model attention. To better understand *what* the model relies on in the background, we visualize Class Activation Maps (CAMs) [65] to see which contextual objects the model's attention focuses on. CAMs are saliency maps shown to expose the implicit attention of neural networks, highlighting "the most informative image regions relevant to the predicted class." [65] For example, the model tends to focus on various spurious correlations such as indoor objects (`oven` and `bed`) to classify female and outdoor objects (`skateboard` and `motorcycle`) to classify male (Fig. 7).

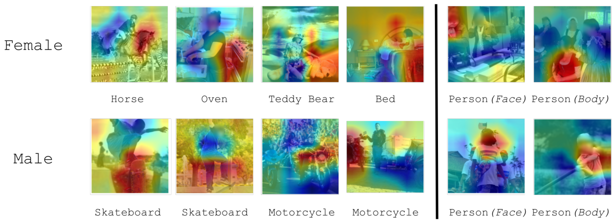

Figure 7. **Visualizing model attention.** On the left four images, the gender artifact model's attention is on contextual objects and not the person, suggesting the model relies on spurious correlations to infer gender. On the right, the model's attention is on the person. These qualitative CAM analyses suggest gender artifacts are embedded in the background of images (i.e., beyond the person) as observed by Hendricks *et al.* [31] and motivate future analysis in contextual objects.

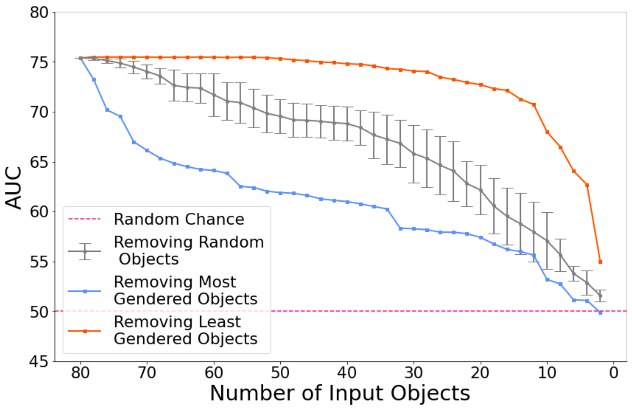

Figure 8. **Performance of contextual objects in the Logistic Regression Model (COCO).** We visualize the change in AUC as objects are iteratively removed from the object-based logistic regression classifier to see how many objects are required before the classifier performs at random chance. When removing random objects, we train five separate random classifiers and report the standard deviation. We report the most relevant objects in descending order, as identified by the weights of a logistic regression classifier trained for gender prediction.

Performance of contextual objects in the logistic regression model. In addition to reporting the results of bootstrapping in the main paper, we also report the results of running five separate random classifiers and report the standard deviation for the model where we remove random objects (Fig. 8).

	COCO		OpenImages	
	B	W	B	W
Full	93.4 ±0.2	93.4 ±0.2	81.1 ±0.3	81.1 ±0.3
Full NoBg	92.7 ±0.2	93.0 ±0.2	-	-
MaskSegm	79.6 ±0.3	78.4 ±0.3	-	-
MaskRect	70.8 ±0.3	70.7 ±0.3	63.1 ±0.4	63.5 ±0.4
MaskSegm NoBg	74.8 ±0.3	76.3 ±0.3	-	-
MaskRect NoBg	58.0 ±0.4	58.3 ±0.4	62.2 ±0.4	58.7 ±0.4

Table 5. **Performance of gender artifact model on various occlusions.** We report the AUC (%) of the gender artifact model on various occlusions when we use black pixels (B) versus white pixels (W) for both COCO [42] and OpenImages [40].

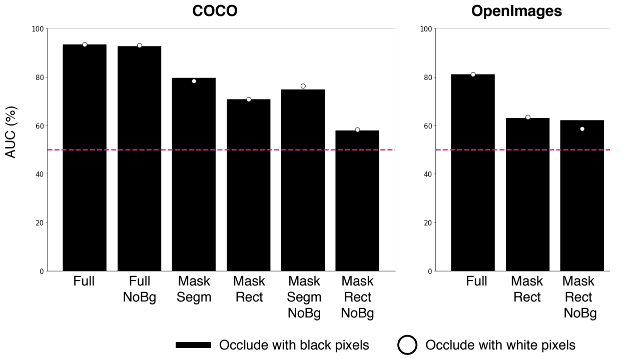

Figure 9. **Comparing AUCs when occluding with different color pixels.** We report the AUC (%) of the gender artifact model when we occlude using black pixels (also reported in Sec. 5) in the black bars and when we occlude using white pixels in the white points. The pink line represent the AUC at random chance.

E. Person and background occlusions

Robustness to occlusion color. In Sec. 5, we occlude person and background cues using black pixels. To demonstrate that occlusions are robust to the color of the pixels, we present the AUCs when we occlude using white pixels instead (see Fig. 9). As seen in Tab. 5, the model performance does not change considerably when we use white pixels as opposed to black pixels for our occlusions. The largest change in AUC for COCO is 1.5% on MaskSegm NoBg and for OpenImages is 3.5% for MaskRect NoBg. Further, the ranking of AUCs across occlusions does not change.

Additional experimental settings. In addition to the six settings we present in Sec. 5, we analyze three more settings. In the first, we occlude the background around the bounding box of the person. This achieves an AUC of 93.0 ± 0.2 and 89.6 ± 0.2 for COCO and OpenImages respectively. Next, we occlude only the person's face which was detected using Amazon Rekognition and MTCNN [81]. This yields an AUC of 92.0 ± 0.2 and 79.8 ± 0.3. Finally, only on the COCO images, we occlude the background and include only the person's skeleton, which achieves an AUC of 65.6 ± 0.4.

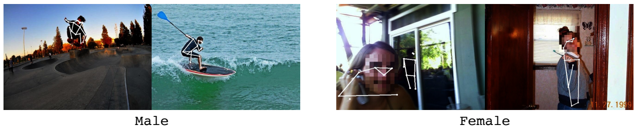

Figure 10. **Qualitative examples of differences in poses.** We visualize poses, overlayed on the original COCO image, that were predicted as highly male (top) and highly female (bottom). Face pixelation is not in the COCO image but included in an effort to partially preserve privacy.

	Original	Balanced
Pix- Sh-Bg	93.4	93.3
Pix-Sh-SomeBg	93.0	93.0
Pix-Sh-NoBg	92.6	92.6
NoFace-Sh-Bg	92.0	91.7
NoPix-Sh-Bg	79.4	78.1
NoPix-NoSh-Bg	70.7	69.4
NoPix-Sh-NoBg	76.0	76.7
NoPix-SomeSh-NoBg	65.6	63.6
NoPix-NoSh-NoBg	58.4	58.9

Table 6. **Results on gender-balanced datasets** We report the AUC (%) of the model trained on a gender-balanced dataset

Pose analysis. After training a model just on the person's keypoints from COCO, we qualitatively inspect the poses with the greatest absolute scores. In Fig. 4, we display the poses predicted to be more likely to be male and female and notice that images predicted to. be male are smaller and in action (e.g., playing a sport, jumping) whereas those predicted to be more likely to be female tended to be larger and standing still.

F. Results on a balanced dataset

Both COCO and OpenImages are skewed male (69.2% and 61.1% of the training set for COCO and OpenImages). We examine whether the gender imbalance in the dataset affects the discoverable artifacts. Concretely, we train and evaluate our gender artifact model on a balanced dataset. As shown in Tbl. 6, the AUCs are comparable to the unbalanced dataset. In fact, the largest change of AUC was from 65.6 on the skewed dataset to 63.6 on the balanced for NoPix-SomeSh-NoBg.

G. Additional fairness-through-blindness results

We provide additional results on the COCO objects that were most likely to be affected by the adversarial de-biasing method. The top five non-`person` objects, which we calculate as the ratio of images in which the object is affected to the number of images that object occurs in, are as follows: `giraffe`, `train`, `elephant`, `bus`, and `horse`. It is likely that these objects are more likely to be occluded as a person may be directly interacting with the object (e.g., riding a `horse`) in the image. We provide the full COCO object results in Fig. 11.

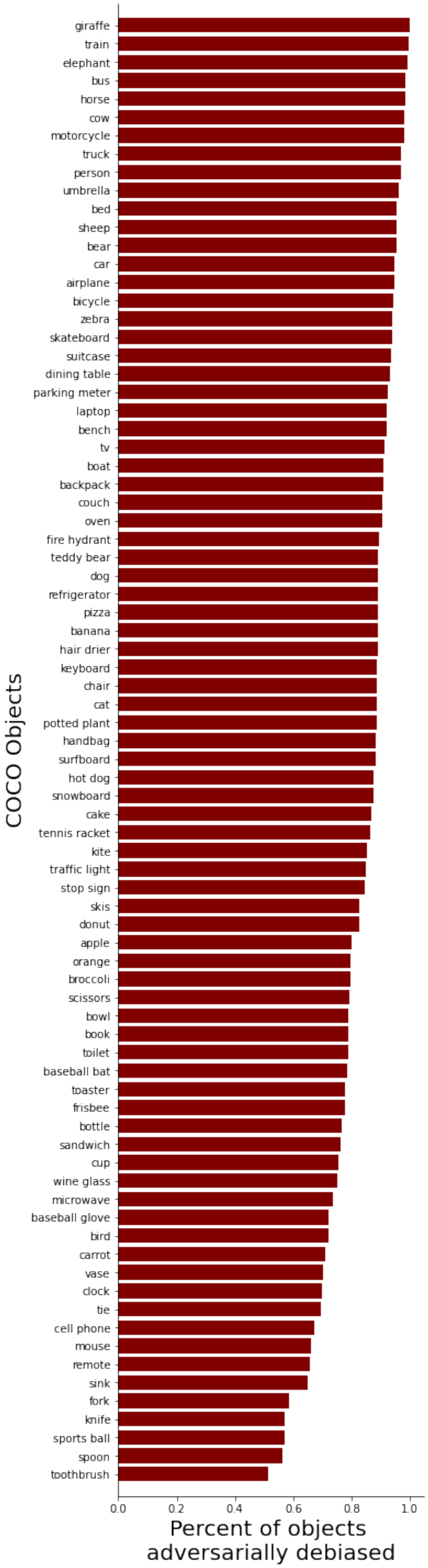

Figure 11. **Occluded contextual objects during adversarial debiasing.** We visualize the additional results from our fairness-through-blindness case study in which we analyze the COCO objects that were most likely adversarially debiased (i.e., contribute to gender information).