Appendix

Table of Contents

A. Full Training Settings

In this section we provide the complete details regarding the training setting for our dense and sparse models on CelebA. For all our experiments we used standard random augmentations for CelebA used in [50], and we normalized the samples using mean and standard deviation each of 0.5 per channel. Furthermore, we replicated all experiments from five different seeds. We adapted the public implementation for model pruning: https://github.com/IST-DASLab/ACDC to train with Binary Logistic Loss.

Joint training. We train the dense model for 100 epochs, using SGD with momentum, with the same hyperparameters (learning rate scheduler, momentum, weight decay, batch size) as the ones used for training ImgageNet in [37], but without label smoothing. Generally, we have noticed that on the held-out CelebA validation set, the dense model tends to overfit after around 40 epochs; therefore, we consider the model with the best validation during training and we use it for our final results on the test set. Likewise, we use the same training hyperparameters for GMP-RI; furthermore, we start pruning from the 10th epoch, using global magnitude pruning on all layers, and increase the sparsity level every 10 epochs, using a standard polynomial schedule [55]. We finetune the sparse models for the last 20 epochs of training and consider the models with the best validation between epochs 80-100. In the case of GMP-PT models, we use 80 epochs for training, and we increase the sparsity level every 4th epoch, while the final 20 epochs are reserved for finetuning at maximum sparsity. For GMP-PT we use the Adam optimizer, with a fixed learning rate of 0.0001, similar to [39].

Single label training. In addition to the joint attribute training, we also train a subset of labels individually. The labels we consider are the following: Bags Under Eyes, Blond, Big nose, Mustache, Oval Face, Receding Hairline, and Smiling. All single label experiments are trained for 20 epochs to avoid overfitting. The dense models were trained using SGD with momentum, with initial learning rate 0.1, batch size 256, momentum value 0.9 and weight decay 0.0001; additionally, we used a cosine annleaning learning rate scheduler. The GMP-RI models were trained using SGD with momentum value 0.9, weight decay 0.0001 and fixed learning rate of 0.1; models were pruned starting from the third epoch, with a gradual increase in sparsity every epoch following a polynomial schedule [55], while the final 4 epochs were reserved for finetuning.

B. Full Override Results for Jointly-trained ResNet18 Models

In this section, we present the full data for the impact on Bias Amplification of selectively overriding model predictions with dense predictions (in the case of sparse models) or correct labels. In all cases, the overridden samples are prioritized by the uncertainty of the *dense* model on that attribute. Further, only predictions for attributes that show positive bias amplification in the dense case are overridden. The results are shown in Figure B.1. We observe that in general, overrides using dense model predictions are effective in the case of very sparse (99%-99.5% sparse) models, but their effectiveness decreases for less sparse models. This is consistent with our observation that less sparse models show less bias amplification relative to dense even without any interventions. Further, we observe that even for categories where the BA is relatively low (Chubby and Pale Skin), overrides are still effective at further reducing relative bias amplification at high sparsity. Overriding with the true label reduces bias amplification throughout.

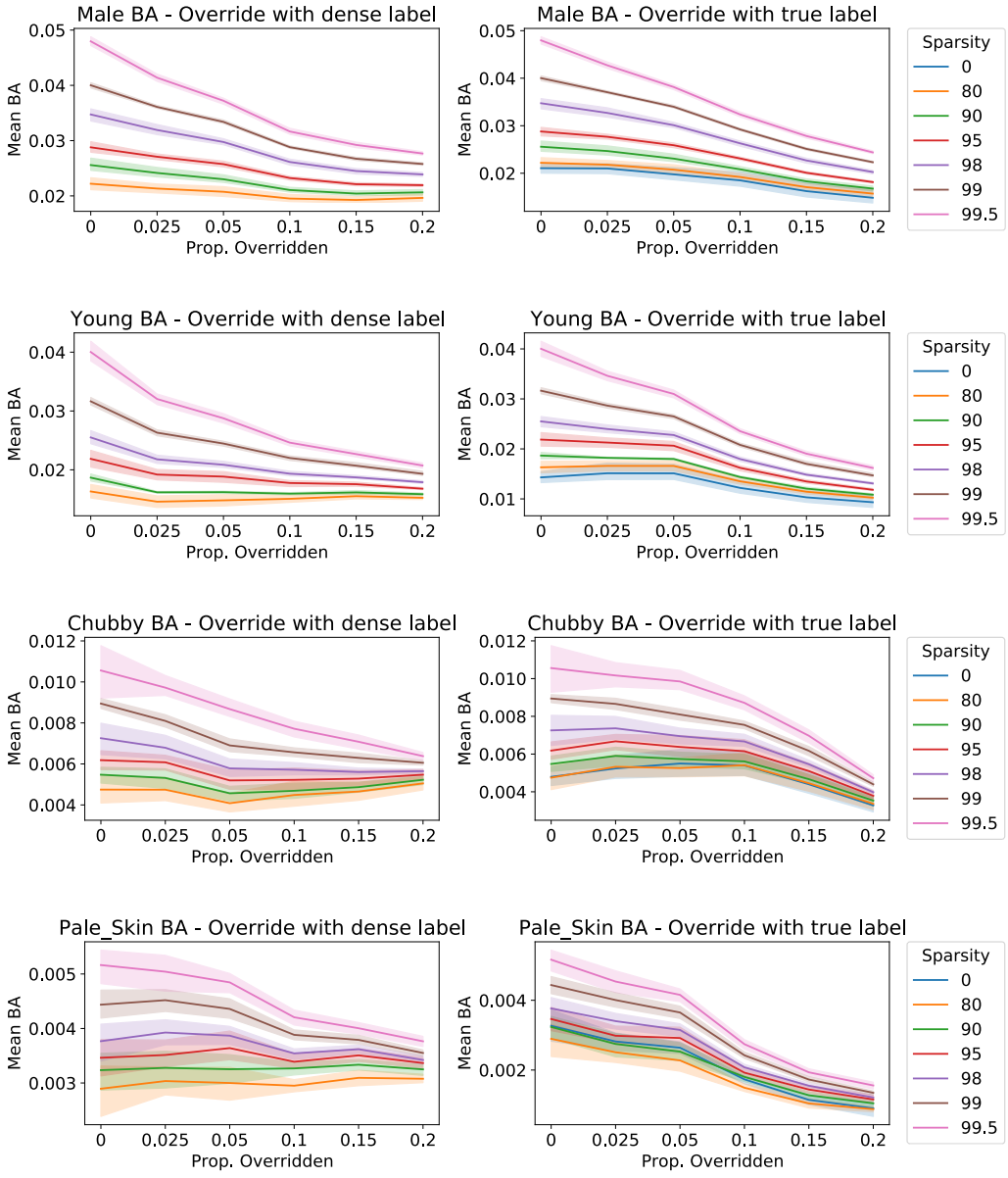

Figure B.1. [CelebA / ResNet18 / GMP-RI] Effect of label overrides on Bias Amplification. In all cases, overrides are prioritized by dense model uncertainty.

C. Full results for Singly-trained CelebA models on ResNet18

In this section we provide and discuss Figure C.2, which is a more complete version of Figure 3 (Accuracy and Bias on singly-trained models); this version includes all seven binary attributes for which we ran the experiment, and all metrics. We observe that the conclusions which we described in Sections 3 for the Oval Face and Big Nose attributes generally hold true for the additional five attributes (Bgs Under Eyes, Receding Hairline, Mustache, Blond Hair, and Smiling) as well. We observe that model accuracy and AUC is generally higher for single-attribute models than joint models, at no or low sparsities, but roughly equal for high sparsities. Further, singly-trained models are much less impacted by sparsity than jointly-trained models when it comes to both Systematic and Categorical bias. However, this manifests as *less* bias in jointly-trained models at low sparsity, and roughly equal bias at high sparsities ($\geq 95\%$).

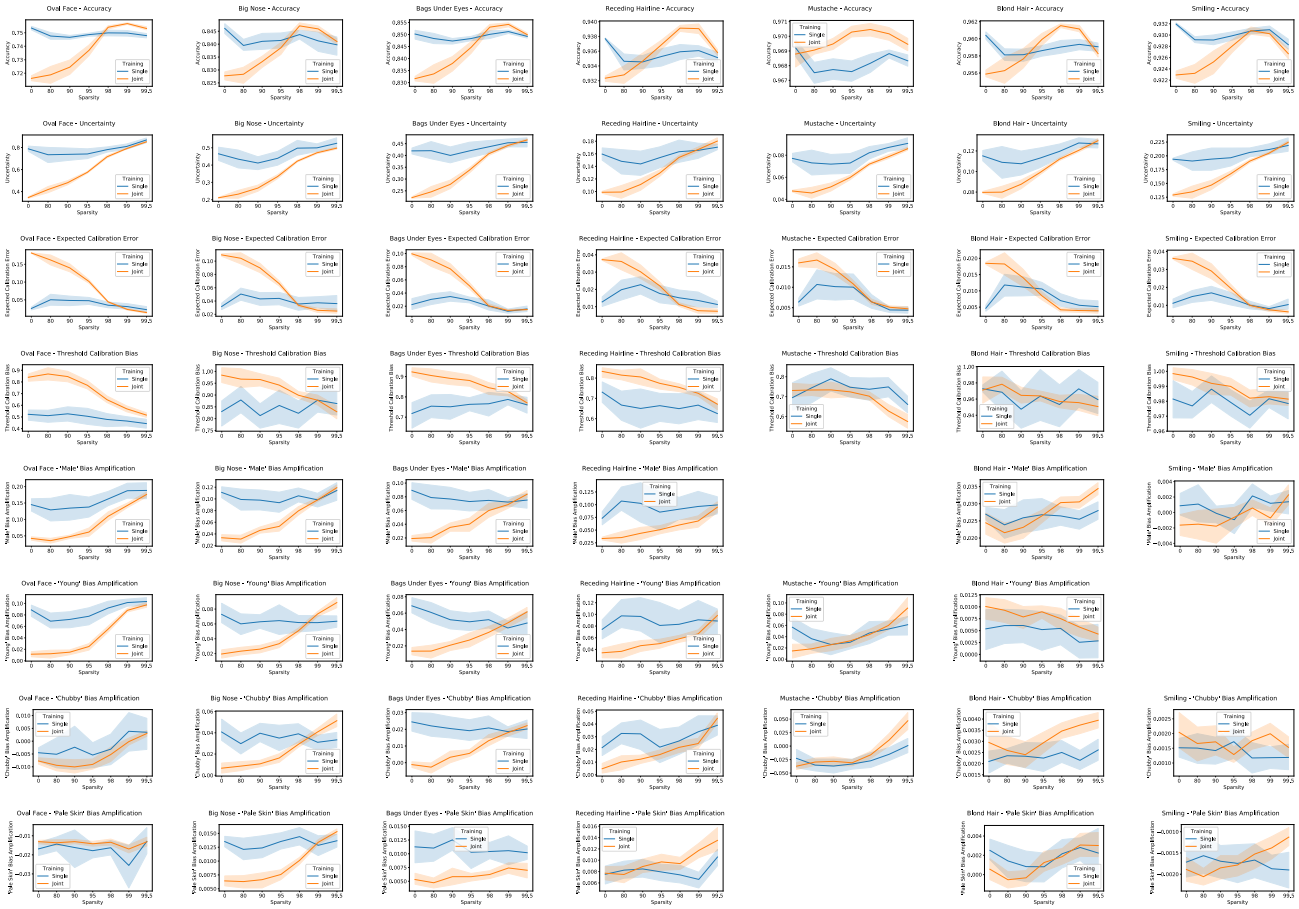

Figure C.2. [CelebA / ResNet18 / Single Attribute / GMP-RI] Effect of single versus joint training of attributes on Accuracy (first row), Uncertainty (second row), ECE (third row), Threshold Calibration Bias (fourth row), and Bias Amplification for the 'Male', 'Young', 'Chubby', and 'Pale Skin' attributes (fifth-eighth rows), on the ResNet18 CelebA model, predicting, from left to right, Oval Face, Big Nose, Bags Under Eyes, Receding Hairline, Mustache, Blond Hair, and Smiling). Orange denotes results from joint runs and blue denotes results from single runs. Omitted panels are cases where BA cannot be computed, either because there is no relationship between the predicted attribute and the category, or because the attribute is not present for one of the values of the category.

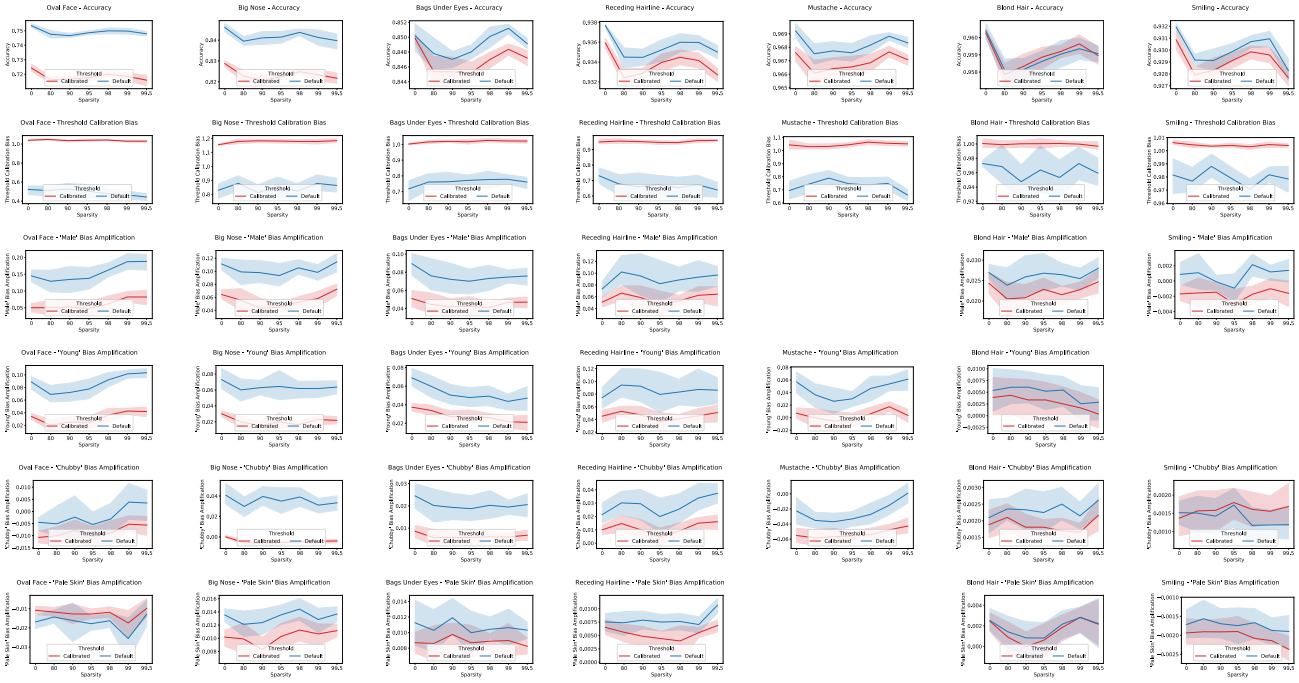

Figure C.3. [CelebA / ResNet18 / Single Attribute / GMP-RI] Effect of threshold adjustment on Accuracy (first row), Threshold Calibration Bias (second row), and Bias Amplification for the 'Male', 'Young', 'Chubby', and 'Pale Skin' attributes (third-sixth rows), on the ResNet18 CelebA model, predicting, from left to right, Oval Face, Big Nose, Bags Under Eyes, Receding Hairline, Mustache, Blond Hair, and Smiling). Red denotes results where the threshold is calibrated on the validation set, and blue denotes results from runs where the default threshold of 0.5 was used. Omitted panels are cases where BA cannot be computed, either because there is no relationship between the predicted attribute and the category, or because the attribute is not present for one of the values of the category.

D. Bias Amplification Results from Training the Predicted and Category Attribute Together

Inspired by our observation that, at low sparsities, joint training of all 40 attributes results in substantially lower bias amplification, we tested the impact of jointly training two attributes - a predicted attribute that shows high bias amplification in other training scenarios, and the identity category with regard to which high BA was observed. In all, we jointly co-trained five such pairs: Big Nose + Male, Oval Face + Male, Big Nose + Young, Mustache + Young, and Receding Hairline + Young. Except for using two logistic heads in the prediction layer, the training setting matches exactly our training settings for singly-trained models.

The results of the experiment are shown in Figure D.4. We observe that in all five cases, the BA of the ”double” model, which co-trains the protected and predicted attribute, matches the BA of the single model fairly closely. This result suggests that more attributes looking at various facial features would need to be jointly trained in order to decrease BA at lower sparsities.

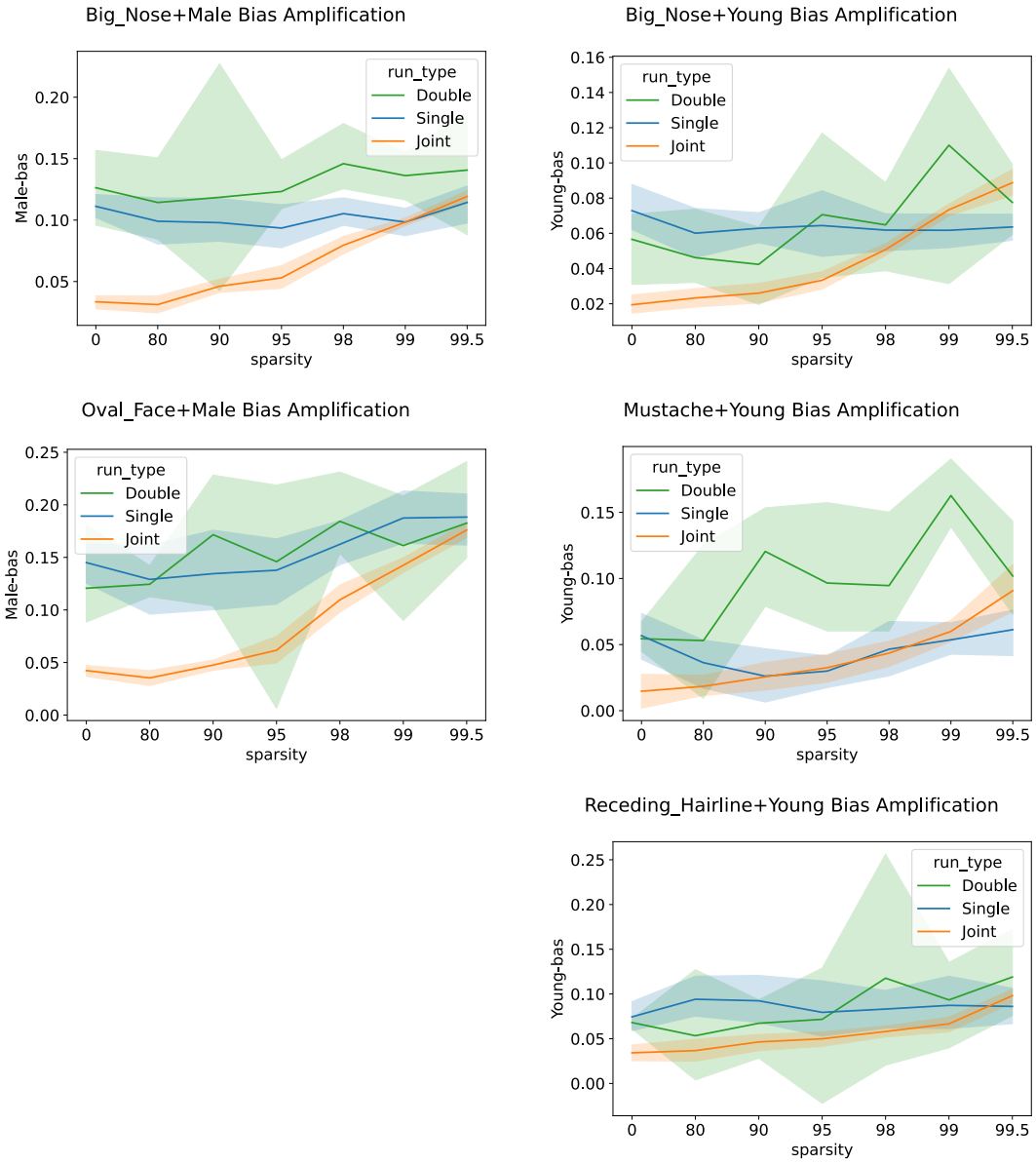

Figure D.4. [CelebA / ResNet18 / Two-Attribute / GMP-RI] Comparison of bias amplification between models that are singly-trained, jointly-trained for all forty attributes, and models that are trained to predict only one attribute + the protected category.

E. Post-training pruning results

We further extend our analysis of bias in sparse CelebA/ResNet18 models, by using a different pruning procedure. Specifically, we perform gradual magnitude pruning starting from pre-trained dense models (GMP-PT); the full training hyperparameters are explained in Appendix Section A. Our results for GMP-PT are presented in Figure E.5. In terms of accuracy or AUC performance, we obtain good quality models even at high sparsity ($> 99\%$), which is in line with our observations for the GMP-RI setting. Similarly, our conclusions hold for Systematic and Category bias. Namely, the ECE and TCB go down with sparsity, while the interdependence slightly increases and the fraction of uncertain samples increases substantially with model sparsity. The Category bias (BA) also increases with sparsity; this can be seen better on the Male attribute. Notably, compared to GMP-RI, the BA values are slightly lower for less sparse models (*e.g.* 80% and 90% sparse). We further test methods for bias mitigation on the GMP-RT and notice similar effects to the GMP-RI setting; namely, when overriding low confidence samples in the sparse models with either the true or dense label, we observe a substantial decrease in Category bias, as measured by BA, particularly at high sparsity (please see Figure E.7). Lastly, we study the relationship between uncertain samples and compression identified exemplars (CIEs) [28, 29] in Figure E.6 and observe that most of the CIEs are uncertain samples.

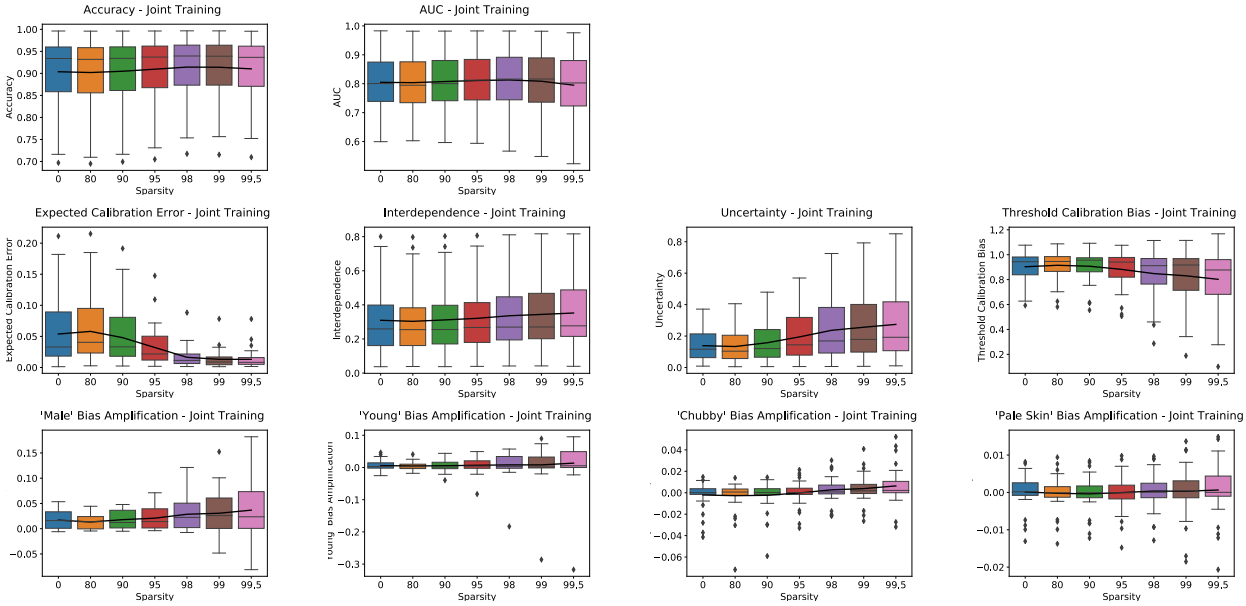

Figure E.5. [CelebA / ResNet18 / GMP-PT] Accuracy and Systematic Bias metrics (TCB, ECE, Interdependence) of ResNet18 models jointly trained on all CelebA attributes, and pruned Post-Training (GMP-PT). The thick black line denotes the mean value at each sparsity level.

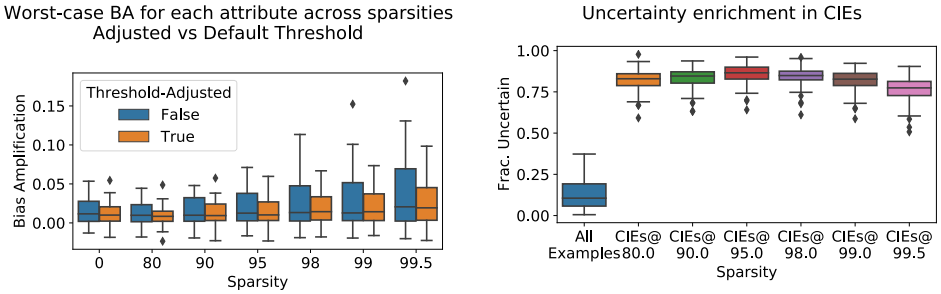

Figure E.6. [CelebA / ResNet18 / GMP-PT] (Left) Effect of threshold calibration on models jointly trained on all attributes. (Right) Proportion of uncertain predictions for *dense* models across all attributes for all elements in the CelebA test set, and for Compression-Identified Exemplars at different sparsities.

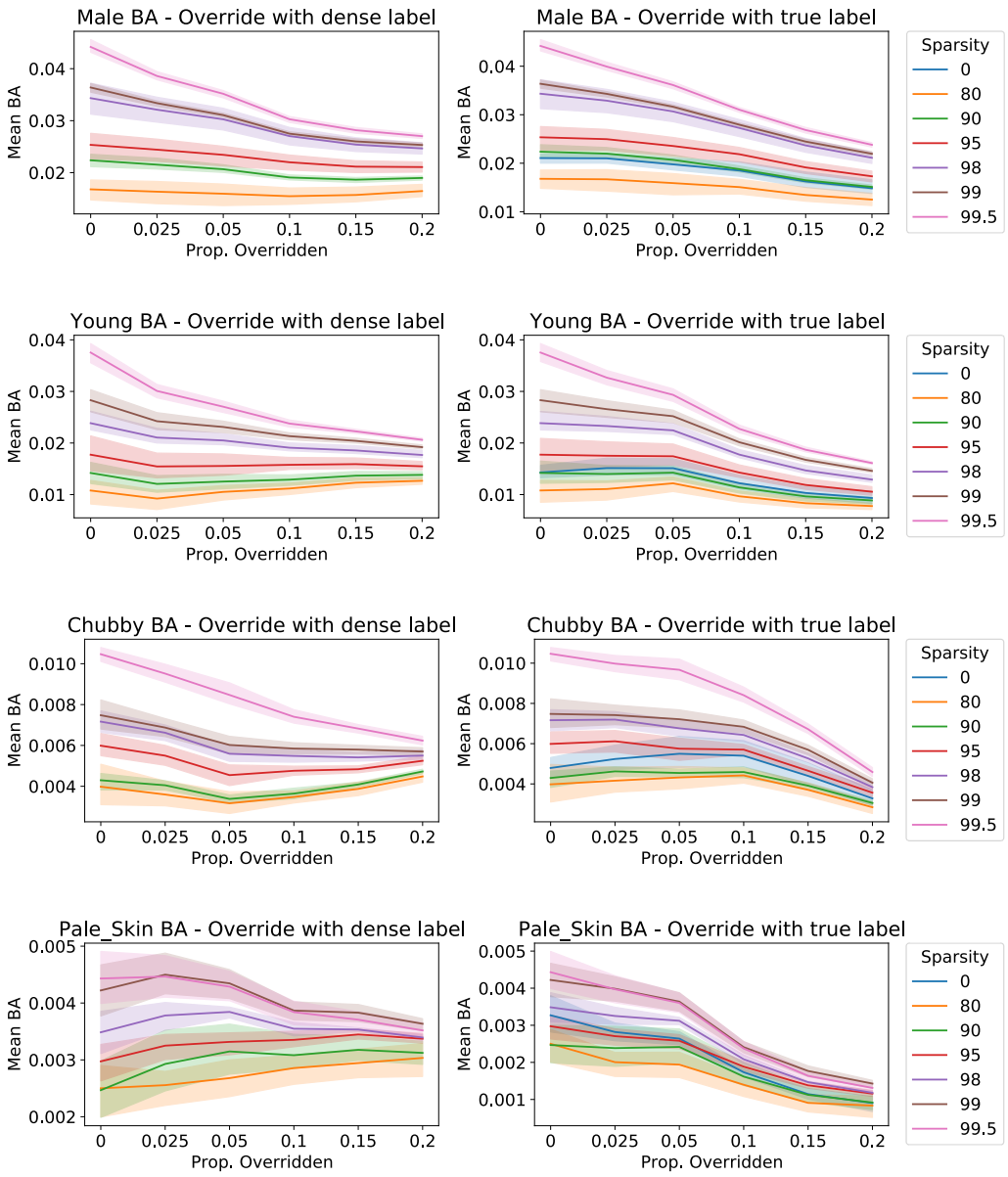

Figure E.7. [CelebA / ResNet18 / GMP-PT] Effect of label overrides on Bias Amplification. In all cases, overrides are prioritized by dense model uncertainty.

F. N:M Sparsity Results

While modern GPU hardware cannot take full advantage of unstructured sparsity, introducing additional constraints can lead to effective speedups. In particular, N:M sparsity patterns, in which N out of every contiguous M values are removed, can be successfully accelerated [44]. We validate our findings by evaluating systematic and categorical bias in the N:M sparsity setting. The sparsification algorithm is a variant of the Random-Initialization Global Magnitude Pruning algorithm used in the main body of the paper. Each experiment was repeated from three different random initializations.

We present our results in Figure F.8. As in our other experiments, we observe little effect on accuracy and AUC even at the highest 1:8 sparsity level; further, we observe that, as with unstructured sparsity, Expected Calibration Error decreases slightly with sparsity, while Uncertainty increases and Threshold Calibration Bias gets slightly worse. As far as Bias Amplification, we observe a slight increase when splitting the data by the Male category, for the 1:4 and 1:8 sparsity pattern. Splitting by the other three categories (Young, Chubby, and Pale Skin) shows minimal, if any, increased BA, likely because even at the highest 1:8 sparsity level, the model is less than 90% sparse, as compared with up to 99.5% sparsity for unstructured pruning. We note that this further validates our finding that ResNet18 models predicting CelebA attributes can be pruned to fairly high sparsity without significant effect on BA.

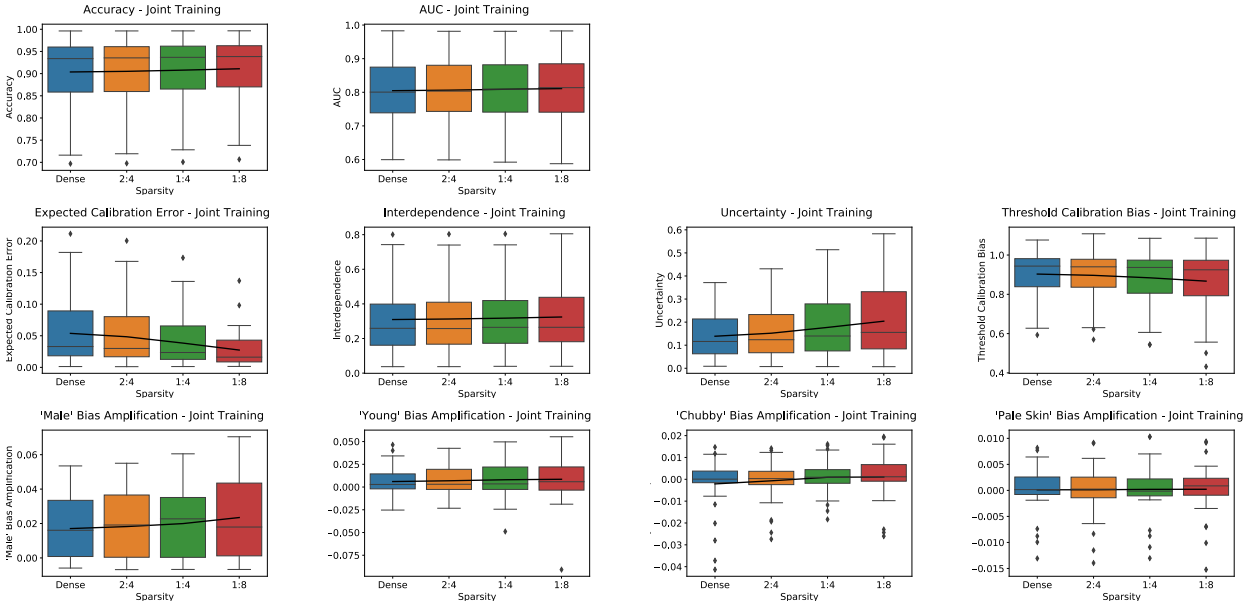

Figure F.8. [CelebA / ResNet18/ N:M/ GMP-RI] Accuracy and Systematic Bias metrics (TCB, ECE, Interdependence) of MobileNetV1 models jointly trained on all CelebA attributes. The thick black line denotes the mean value at each sparsity level.

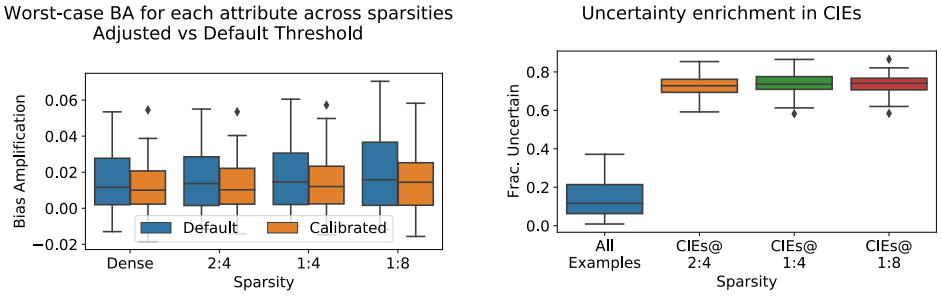

Figure F.9. [CelebA / ResNet18 / N:M Sparsity / GMP-RI] (Left) Effect of threshold calibration on ResNet18 N:M sparsity models jointly trained on all attributes. (Right) Proportion of uncertain predictions for *dense* models across all attributes for all elements in the CelebA test set, and for Compression-Identified Exemplars at different sparsities.

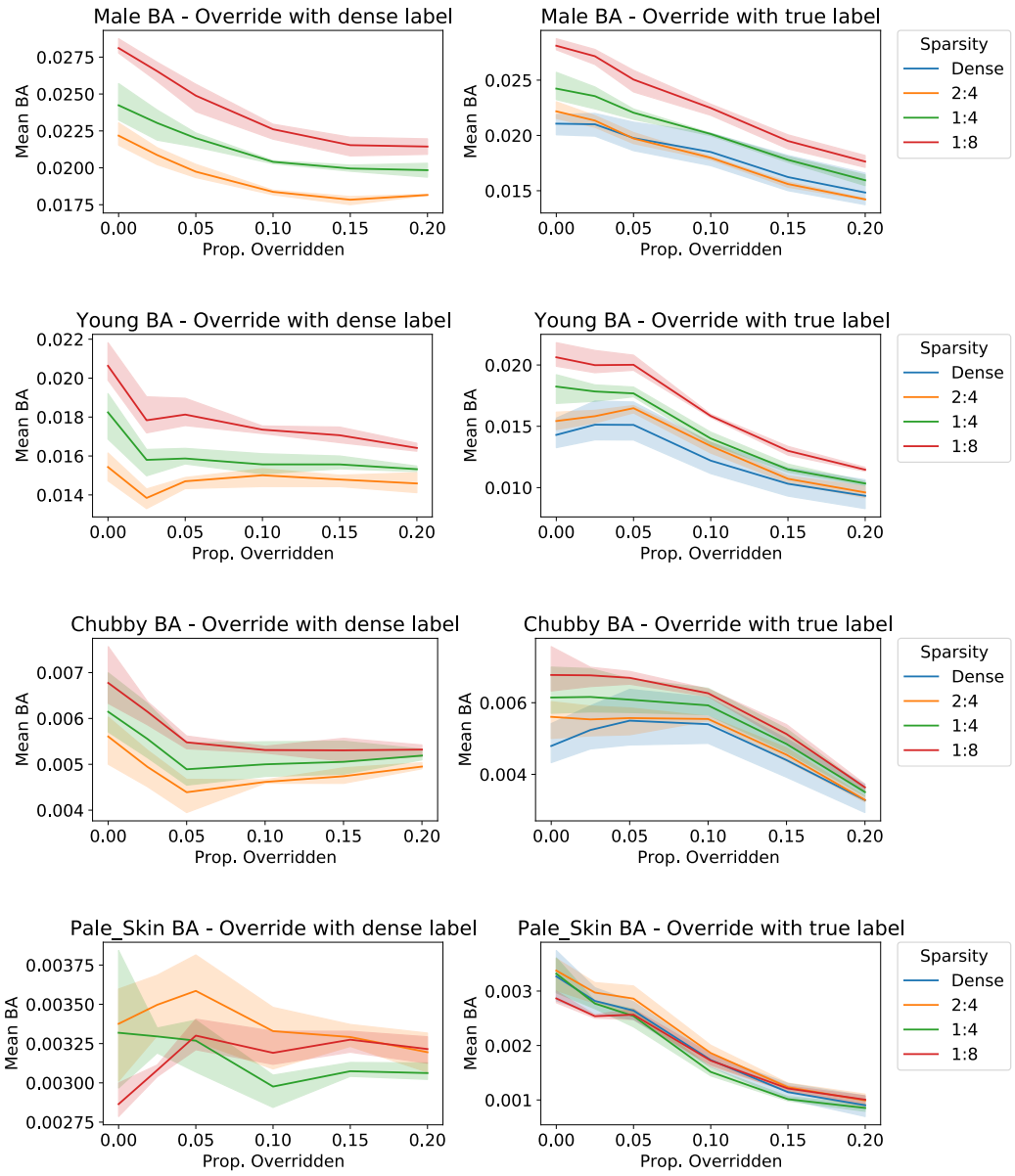

Figure F.10. [CelebA / ResNet18 / N:M Sparsity / GMP-RI] Effect of label overrides on Bias Amplification. In all cases, overrides are prioritized by dense model uncertainty.

G. MobileNetV1 results

We additionally validate our results on different architectures, for the joint label training setting. Namely, we choose MobileNet [30], as it is a smaller model, and known to be more difficult to prune. We train the dense and sparse models using the same hyperparameters described in Appendix Section A. We show results under the GMP-RI setting.

For the MobileNet architecture, we note that sparse models maintain a good performance relative to dense, except for 99% and 99.5% sparsity, where we observe a decrease in performance, both in terms of accuracy and AUC scores (the 99.5% models in particular are very poor and are omitted from analysis). The results for systematic and context bias in Figure G.11 show similar trends to those observed for ResNet18; we note that all our bias metrics, including uncertainty, are substantially amplified at 99% sparsity, which is not surprising given the lower performance of the model. Moreover, we show in Figure G.13 that it is possible to decrease the bias in 99% sparse models by over-ridding the labels of the low confidence samples with their true or dense labels, and we also show that most of CIEs are uncertain samples in Figure G.12.

We also repeat the single-label experiments on this architecture. Unlike the joint training, performance on singly-trained MobileNet models does not decrease at the 99% sparsity level, which can be observed in Figure G.14. Generaly, we observe similar trends in both Systematic and Categorical bias as we observe on ResNet18.

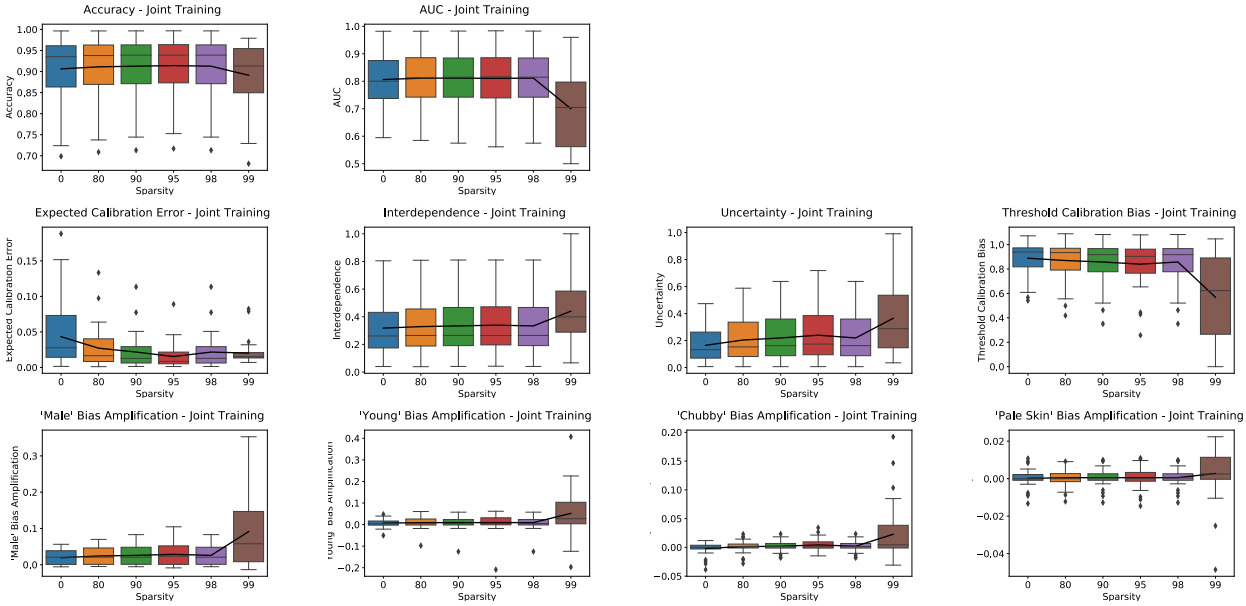

Figure G.11. [CelebA / MobileNetV1 / GMP-RI] Accuracy and Systematic Bias metrics (TCB, ECE, Interdependence) of MobileNetV1 models jointly trained on all CelebA attributes. The thick black line denotes the mean value at each sparsity level.

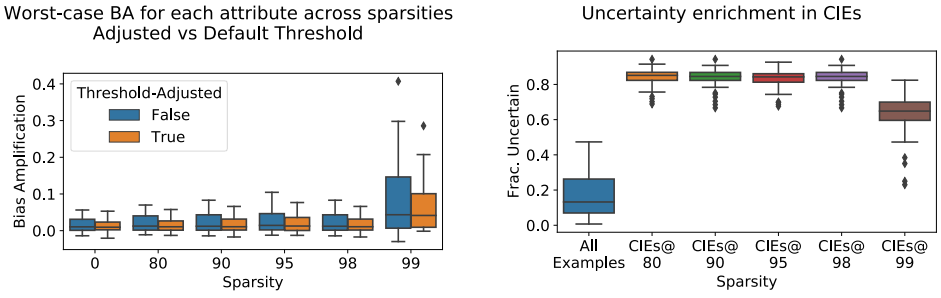

Figure G.12. [CelebA / MobileNetV1 / GMP-RI] (Left) Effect of threshold calibration on models jointly trained on all attributes. (Right) Proportion of uncertain predictions for *dense* models across all attributes for all elements in the CelebA test set, and for Compression-Identified Exemplars at different sparsities.

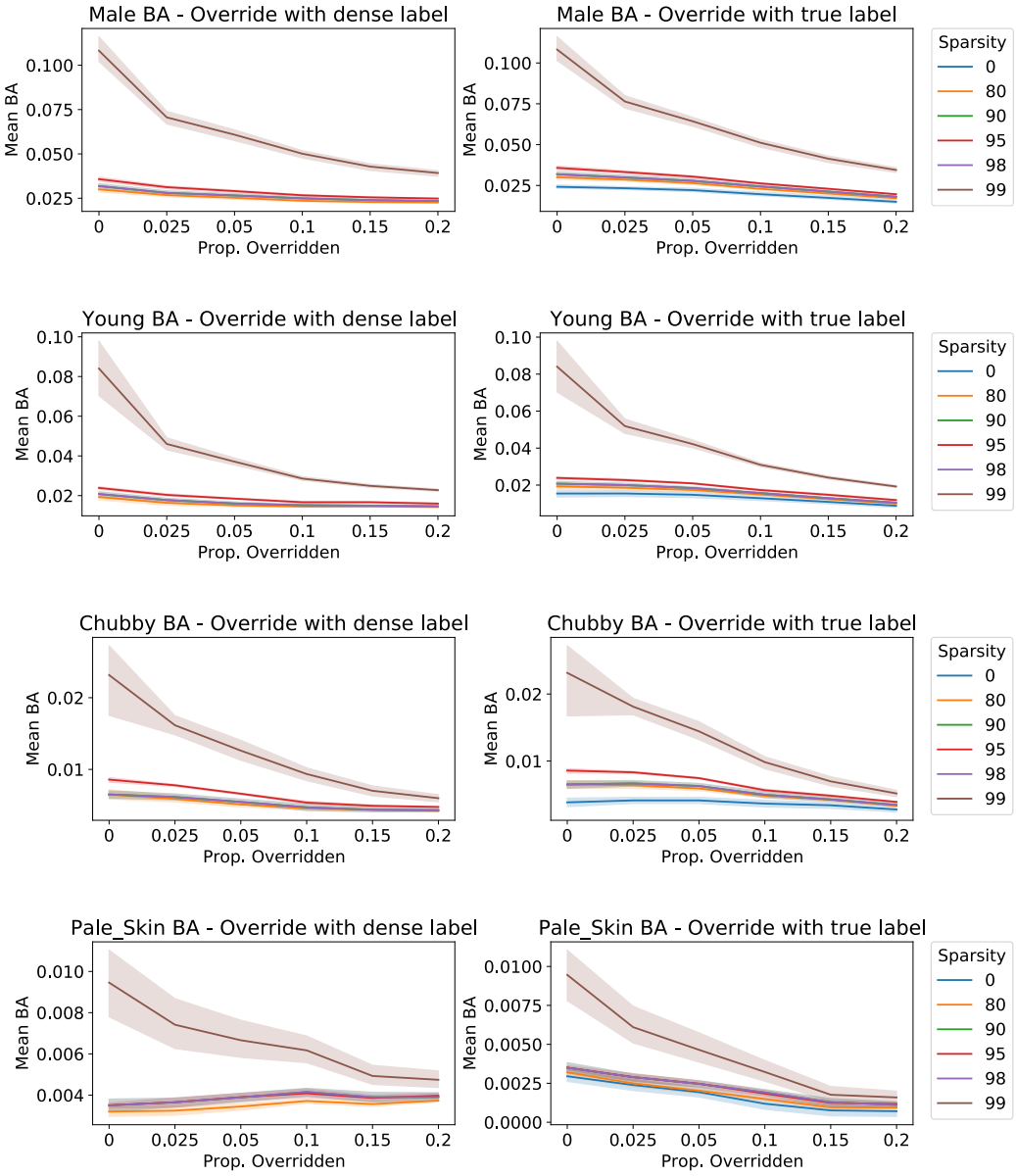

Figure G.13. [CelebA / MobileNetV1 / GMP-RI] Effect of label overrides on Bias Amplification. In all cases, overrides are prioritized by dense model uncertainty.

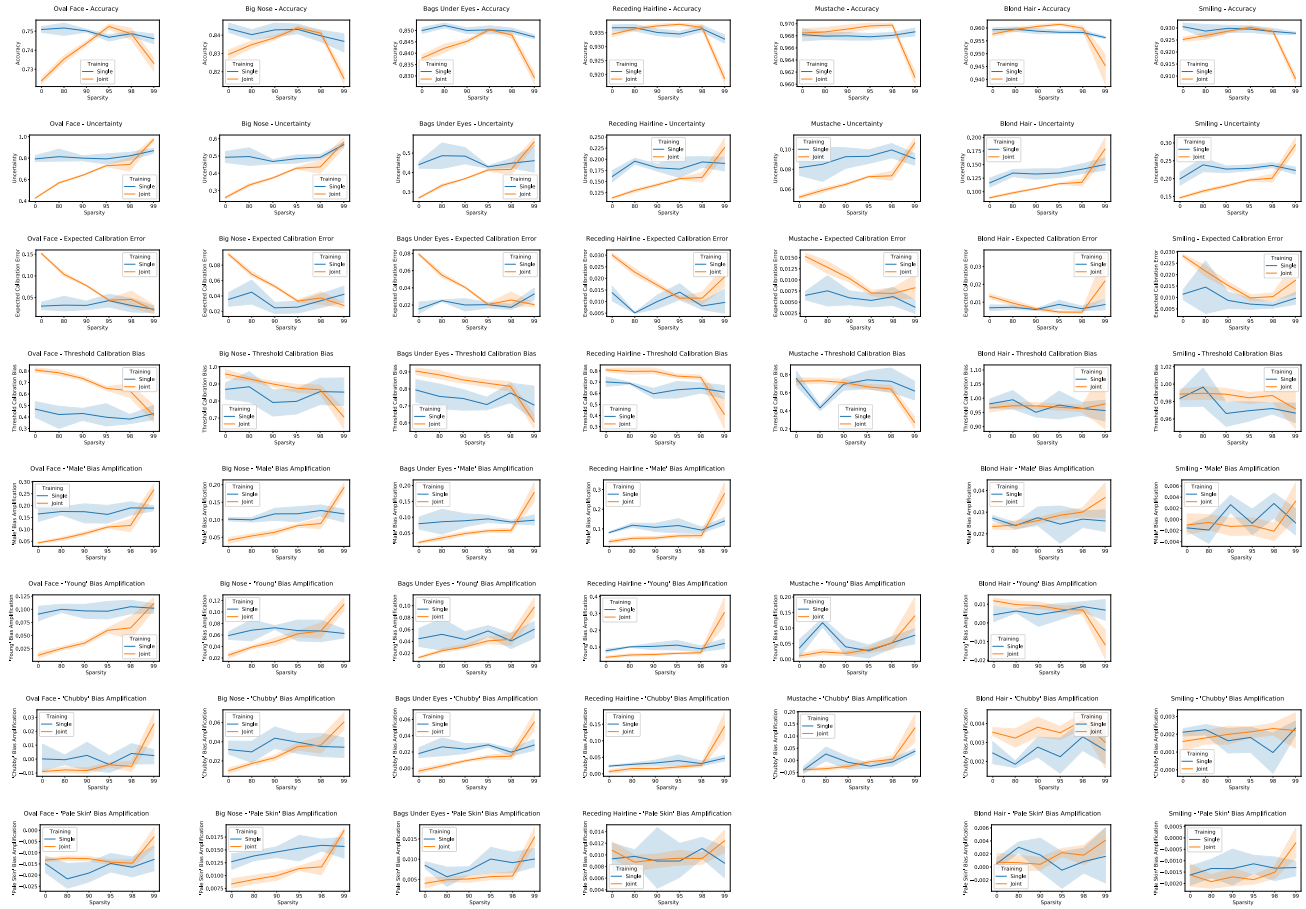

Figure G.14. [CelebA / MobleNetV1 / Single Attribute / GMP-RI] Effect of single versus joint training of attributes on Accuracy (first row), Uncertainty (second row), ECE (third row), Threshold Calibration Bias (fourth row), and Bias Amplification for the 'Male', 'Young', 'Chubby', and 'Pale Skin' attributes (fifth-eighth rows), on the MobileNet CelebA model, predicting, from left to right, Oval Face, Big Nose, Bags Under Eyes, Receding Hairline, Mustache, Blond Hair, and Smiling). Orange denotes results from joint runs and blue denotes results from single runs. Omitted panels are cases where BA cannot be computed, either because there is no relationship between the predicted attribute and the category, or because the attribute is not present for one of the values of the category.

H. ResNet50 Results

We further validate our joint training GMP-RI results on the ResNet50 architecture, which has roughly double the parameters of ResNet18 (25.529.472 versus 11.683.712). We use the same experimental settings as for the ResNet18 GMP-RI experiments, excepting that the ResNet50 experiments were performed only in triplicate (from three random seeds).

The accuracy and systematic bias metrics are presented in Figure H.15. Overall, the patters we observe using the ResNet50 architecture very closely match those using ResNet18. Figure H.17 shows the impact on Bias Amplification of overriding the most uncertain predictions (closest to 0.5 probability as measured on a dense model) with either the dense prediction or the correct label. Consistent with the rest of the paper, the override is only applied if the Bias Amplification is positive on the dense model for the attribute and category in question. As in other cases, both types of overrides are effective at reducing Bias Amplification, generally when using the correct label, and when applied to high-sparsity models in the case of the dense label.

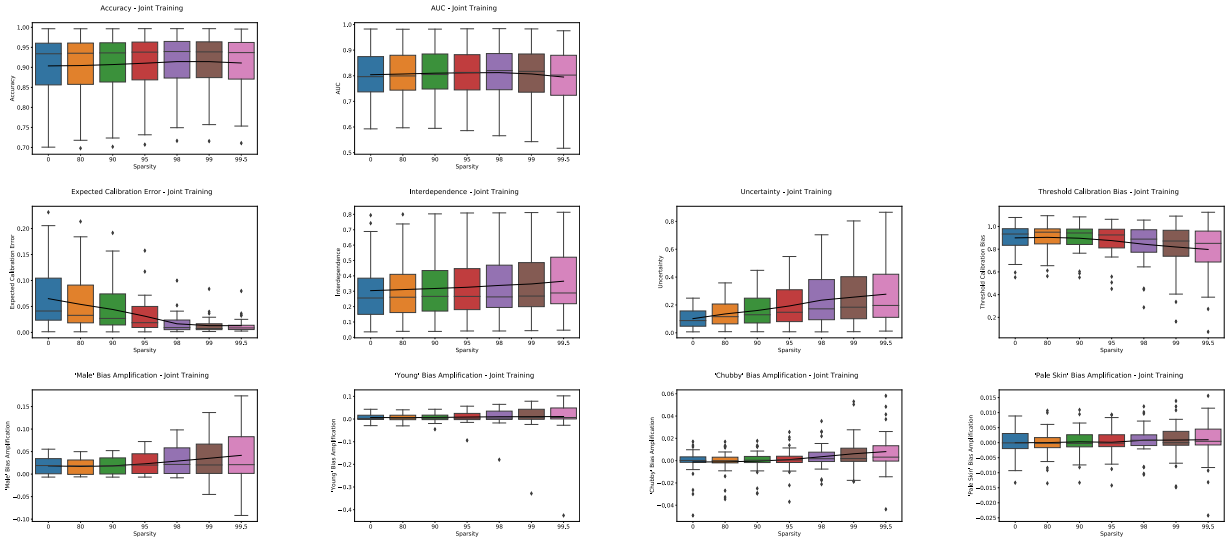

Figure H.15. [CelebA / ResNet50 / GMP-RI] Accuracy and Systematic Bias metrics (TCB, ECE, Interdependence) of ResNet50 models jointly trained on all CelebA attributes. The thick black line denotes the mean value at each sparsity level.

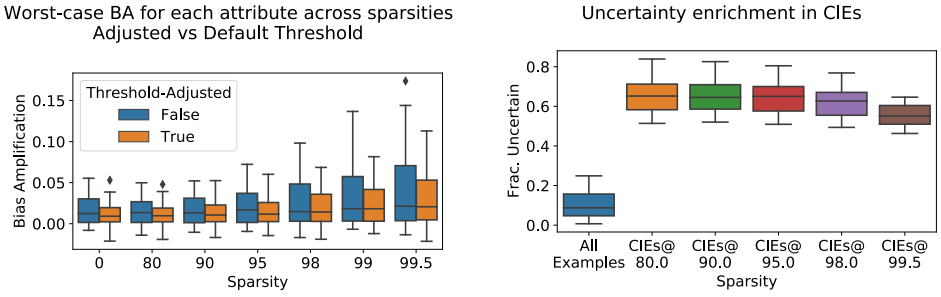

Figure H.16. [CelebA / ResNet50 / GMP-RI](Left) Effect of threshold calibration on ResNet50 models jointly trained on all attributes. (Right) Proportion of uncertain predictions for *dense* models across all attributes for all elements in the CelebA test set, and for Compression-Identified Exemplars at different sparsities.

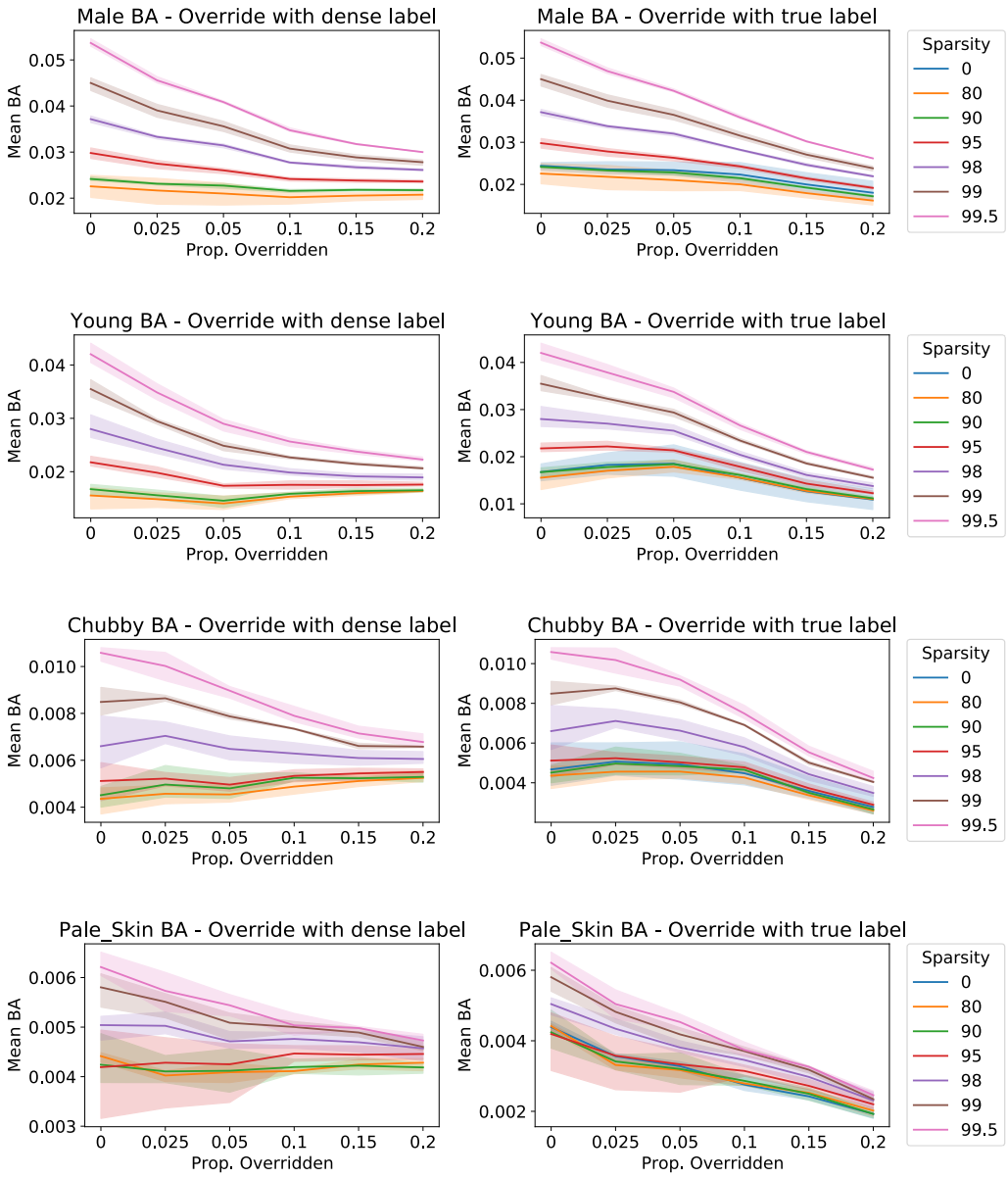

Figure H.17. [CelebA / ResNet50 / GMP-RI] Effect of label overrides on Bias Amplification. In all cases, overrides are prioritized by dense model uncertainty.

I. Uncropped CelebA Results

While inspecting the CelebA samples using our visualization tool described in Appendix Section M we observed that some of the attributes were more prone to mislabelling, due to decisions conventionally made when training models on CelebA; for example, due to the cropping of the images in the standard CelebA version used in practice, it is often times impossible to directly observe the presence of attributes like Wearing Necktie or Wearing Necklace (see the discussion in M, and specifically Figures M.32, M.30). In an effort to disentangle the data inherent bias, due to cropping, from Systematic or Categorical bias, we further validate our results on dense and sparse models trained on the *uncropped* version of CelebA. We use the same setting for training ResNet18 GMP-RI models, as the one described in Appendix Section A. In terms of accuracy or AUC scores, we observe a decrease in performance for very sparse (99.5% sparse) models trained on the uncropped CelebA. Otherwise, our findings in terms of systematic (ECE, TCB, Interdependence) or context (BA) bias generally confirm those on the standard CelebA dataset. It is worth noting, however, that using the uncropped CelebA version substantially reduced the Categorical bias for the problematic attributes Wearing Necklace or Wearing Necktie. For example, the BA scores for the dense model changed from 4.6 to 0.9 for Wearing Necktie and from -2.2 to -1.4 for Wearing Necklace. More importantly, the bias decreased substantially for high sparse models; for example, the interval for the BA scores for models in the 98%-99.5% sparsity range changed from [-34.4, -21.3] for the cropped version to [-5.8, -3.4] for uncropped, for the Wearing Necklace attribute. Similarly, the BA score for Wearing Necktie on the 99.5% sparse model dropped from 8.7 to 3.1, and also decreased substantially for lower sparsity levels. These findings confirm our expectations that data inherent bias can play a significant role in the overall bias equation for a model, and improvements can be obtained by carefully taking the data bias into account. We further show that Categorical bias can be decreased by careful relabelling in Figure I.20 and show the uncertainty of CIEs in Figure I.19.

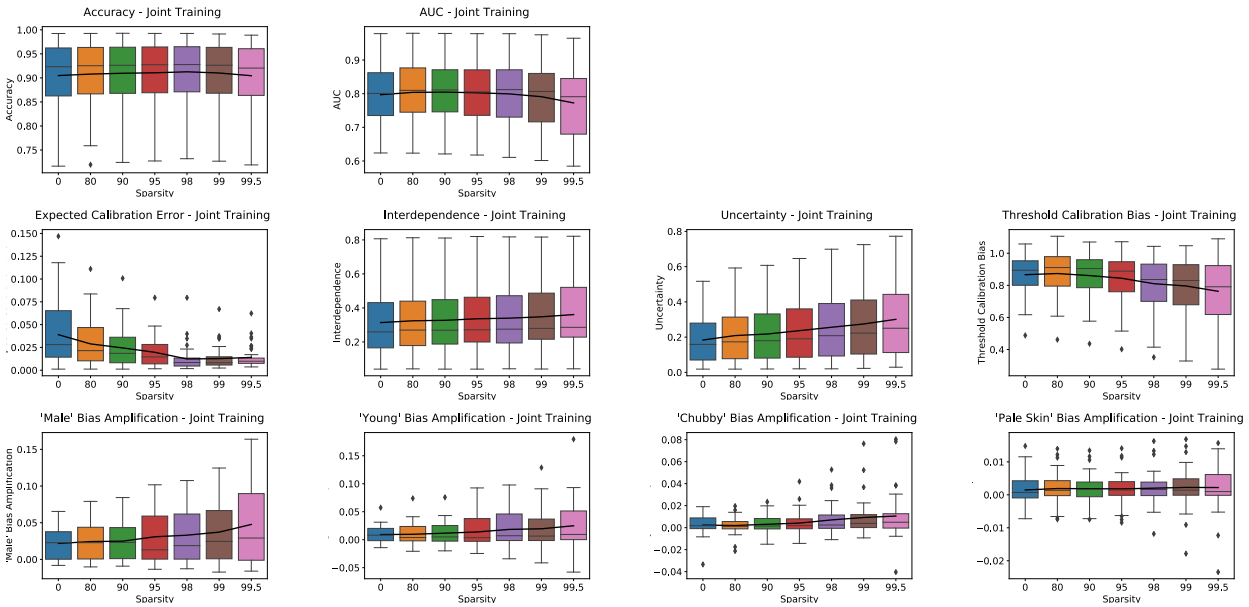

Figure I.18. [Uncropped CelebA / ResNet18 / GMP-RI] Accuracy and Systematic Bias metrics (TCB, ECE, Interdependence) of ResNet18 models jointly trained on all CelebA attributes, using the *uncropped* images for training and inference. The thick black line denotes the mean value at each sparsity level.

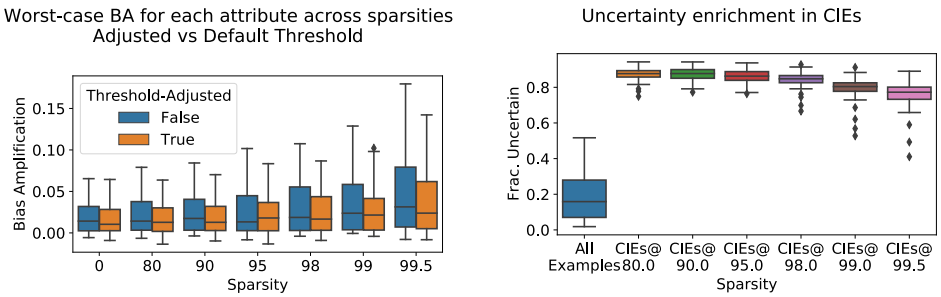

Figure I.19. [Uncropped CelebA / ResNet18 / GMP-RI] (Left) Effect of threshold calibration on models jointly trained on all attributes. (Right) Proportion of uncertain predictions for *dense* models across all attributes for all elements in the CelebA test set, and for Compression-Identified Exemplars at different sparsities.

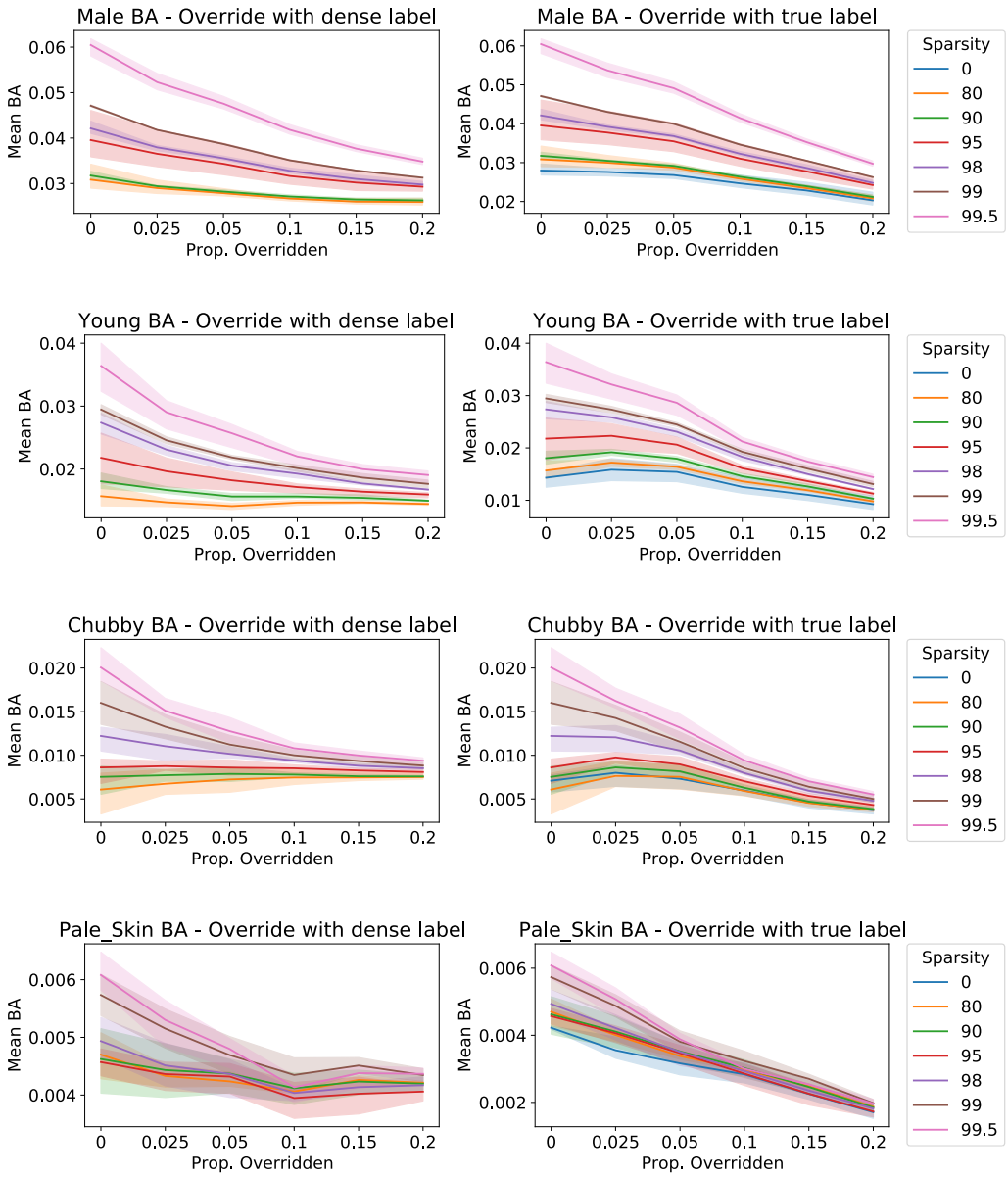

Figure I.20. [Uncropped CelebA / ResNet18/ GMP-RI] Effect of label overrides on Bias Amplification. In all cases, overrides are prioritized by dense model uncertainty.

J. Tabular Results for Jointly-Trained ResNet18 CelebA Models

In this section, we present our main results for systematic and categorical bias metrics for ResNet18 CelebA models in tabular form. We first present the average values across attributes for all metrics by sparsity in Table J.1, then give detailed per-attribute numbers for each metric in subsequent tables. The means and standard deviations were computed from runs from five random seeds.

Table J.1. Mean Accuracy, Systematic Bias, and Categorical Bias Values, Joint CelebA Training, ResNet18

Sparsity Metric	0	80	90	95	98	99	99.5
Accuracy	0.904	0.908	0.909700	0.913	0.915	0.914	0.911
AUC	0.805	0.810	0.813	0.815	0.815	0.810	0.797
Expected Calibration Error	0.0538	0.0401	0.0341	0.0254	0.0153	0.0128	0.0127
Interdependence	0.310	0.319	0.324	0.332	0.341	0.349	0.361
Threshold Calibration Bias	0.903	0.895	0.889	0.877	0.853	0.833	0.805
Uncertainty	0.139	0.172	0.186	0.207	0.237	0.256	0.276
'Male' Bias Amplification	0.0170	0.0180	0.0210	0.0241	0.0294	0.0337	0.0402
'Young' Bias Amplification	0.00600	0.00663	0.00711	0.00851	0.00817	0.0101	0.0148
'Chubby' Bias Amplification	-0.00208	-0.00133	-0.000278	0.00106	0.00269	0.00583	0.00844
'Pale Skin' Bias Amplification	0.000097	-0.000065	0.000323	0.000419	0.000581	0.000645	0.000935

Table J.2. Accuracy, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	94.3 ± 0.1	94.5 ± 0.0	94.6 ± 0.1	94.8 ± 0.1	95.0 ± 0.0	94.9 ± 0.1	94.7 ± 0.1
Arched Eyebrows	82.3 ± 0.1	82.9 ± 0.1	83.2 ± 0.2	83.7 ± 0.2	84.1 ± 0.1	84.0 ± 0.1	83.2 ± 0.1
Attractive	80.5 ± 0.2	81.6 ± 0.1	81.8 ± 0.3	82.6 ± 0.1	83.0 ± 0.1	83.1 ± 0.2	82.7 ± 0.1
Bags Under Eyes	83.2 ± 0.1	83.9 ± 0.2	84.2 ± 0.1	84.8 ± 0.1	85.4 ± 0.2	85.5 ± 0.1	85.0 ± 0.2
Bald	98.9 ± 0.0	98.9 ± 0.0	98.9 ± 0.0	99.0 ± 0.1	99.0 ± 0.0	99.0 ± 0.0	98.8 ± 0.0
Bangs	95.5 ± 0.1	95.8 ± 0.1	95.9 ± 0.0	96.1 ± 0.1	96.2 ± 0.0	96.1 ± 0.0	95.9 ± 0.1
Big Lips	69.7 ± 0.3	70.2 ± 0.1	70.4 ± 0.3	71.3 ± 0.2	72.0 ± 0.1	71.9 ± 0.1	71.3 ± 0.2
Big Nose	82.8 ± 0.2	83.1 ± 0.2	83.6 ± 0.3	84.0 ± 0.3	84.6 ± 0.2	84.6 ± 0.2	84.1 ± 0.3
Black Hair	88.8 ± 0.2	89.3 ± 0.2	89.6 ± 0.1	90.0 ± 0.1	90.3 ± 0.1	90.0 ± 0.1	89.6 ± 0.1
Blond Hair	95.6 ± 0.1	95.9 ± 0.1	95.9 ± 0.1	96.1 ± 0.1	96.2 ± 0.0	96.1 ± 0.1	95.8 ± 0.1
Blurry	96.0 ± 0.1	96.2 ± 0.0	96.2 ± 0.1	96.3 ± 0.1	96.3 ± 0.1	96.3 ± 0.0	96.2 ± 0.1
Brown Hair	87.1 ± 0.2	88.1 ± 0.2	88.6 ± 0.2	89.0 ± 0.1	89.5 ± 0.2	89.3 ± 0.1	89.0 ± 0.1
Bushy Eyebrows	91.9 ± 0.1	92.4 ± 0.1	92.6 ± 0.0	92.9 ± 0.1	93.0 ± 0.1	93.0 ± 0.1	92.8 ± 0.1
Chubby	95.3 ± 0.1	95.5 ± 0.1	95.6 ± 0.1	95.6 ± 0.0	95.7 ± 0.1	95.7 ± 0.0	95.6 ± 0.1
Double Chin	96.0 ± 0.1	96.2 ± 0.1	96.3 ± 0.1	96.3 ± 0.0	96.4 ± 0.1	96.4 ± 0.1	96.2 ± 0.1
Eyeglasses	99.6 ± 0.0	99.6 ± 0.0	99.7 ± 0.0	99.7 ± 0.0	99.7 ± 0.0	99.7 ± 0.0	99.6 ± 0.0
Goatee	97.4 ± 0.0	97.5 ± 0.1	97.5 ± 0.0	97.5 ± 0.1	97.6 ± 0.1	97.5 ± 0.0	97.2 ± 0.1
Gray Hair	98.1 ± 0.1	98.2 ± 0.1	98.2 ± 0.0	98.3 ± 0.1	98.3 ± 0.0	98.3 ± 0.0	98.1 ± 0.0
Heavy Makeup	90.9 ± 0.2	91.2 ± 0.2	91.4 ± 0.1	91.8 ± 0.1	92.0 ± 0.1	91.9 ± 0.1	91.6 ± 0.1
High Cheekbones	86.3 ± 0.2	87.0 ± 0.1	87.2 ± 0.1	87.7 ± 0.1	88.1 ± 0.1	87.9 ± 0.1	87.3 ± 0.1
Male	98.4 ± 0.1	98.4 ± 0.1	98.5 ± 0.1	98.4 ± 0.1	98.4 ± 0.1	98.2 ± 0.1	97.8 ± 0.1
Mouth Slightly Open	93.5 ± 0.1	93.7 ± 0.1	93.9 ± 0.1	94.0 ± 0.1	94.2 ± 0.1	94.1 ± 0.1	93.7 ± 0.1
Mustache	96.9 ± 0.1	96.9 ± 0.1	97.0 ± 0.1	97.0 ± 0.0	97.1 ± 0.1	97.0 ± 0.1	97.0 ± 0.1
Narrow Eyes	86.6 ± 0.2	86.9 ± 0.1	87.0 ± 0.1	87.3 ± 0.1	87.6 ± 0.1	87.7 ± 0.1	87.4 ± 0.0
No Beard	96.0 ± 0.1	96.2 ± 0.1	96.3 ± 0.1	96.4 ± 0.1	96.5 ± 0.1	96.5 ± 0.1	96.2 ± 0.1
Oval Face	71.6 ± 0.2	72.8 ± 0.3	73.3 ± 0.2	74.3 ± 0.1	75.5 ± 0.2	75.7 ± 0.1	75.4 ± 0.3
Pale Skin	97.0 ± 0.1	97.0 ± 0.1	97.2 ± 0.1	97.2 ± 0.1	97.2 ± 0.0	97.2 ± 0.0	97.1 ± 0.1
Pointy Nose	74.3 ± 0.2	75.5 ± 0.2	76.2 ± 0.1	77.0 ± 0.1	77.9 ± 0.1	77.8 ± 0.2	77.2 ± 0.1
Receding Hairline	93.2 ± 0.0	93.5 ± 0.1	93.6 ± 0.1	93.8 ± 0.1	93.9 ± 0.1	94.0 ± 0.1	93.6 ± 0.1
Rosy Cheeks	94.8 ± 0.0	94.9 ± 0.1	95.1 ± 0.1	95.2 ± 0.1	95.2 ± 0.0	95.2 ± 0.1	95.0 ± 0.1
Sideburns	97.8 ± 0.1	97.9 ± 0.1	98.0 ± 0.1	97.9 ± 0.0	98.0 ± 0.0	97.9 ± 0.1	97.7 ± 0.1
Smiling	92.3 ± 0.1	92.5 ± 0.1	92.8 ± 0.1	92.9 ± 0.1	93.1 ± 0.1	93.0 ± 0.1	92.8 ± 0.1
Straight Hair	81.8 ± 0.1	82.8 ± 0.1	83.1 ± 0.2	83.7 ± 0.1	83.9 ± 0.2	83.8 ± 0.1	83.1 ± 0.2
Wavy Hair	81.8 ± 0.1	82.8 ± 0.2	83.3 ± 0.2	83.8 ± 0.2	84.2 ± 0.1	83.9 ± 0.2	83.2 ± 0.1
Wearing Earrings	89.7 ± 0.1	90.2 ± 0.2	90.5 ± 0.1	90.7 ± 0.1	90.9 ± 0.1	90.7 ± 0.1	90.2 ± 0.1
Wearing Hat	99.0 ± 0.1	99.1 ± 0.0	99.1 ± 0.0	99.1 ± 0.0	99.1 ± 0.0	99.0 ± 0.0	99.0 ± 0.0
Wearing Lipstick	93.6 ± 0.1	93.9 ± 0.1	94.0 ± 0.1	94.3 ± 0.1	94.3 ± 0.1	94.4 ± 0.2	94.2 ± 0.1
Wearing Necklace	84.4 ± 0.2	85.2 ± 0.2	85.5 ± 0.1	86.0 ± 0.1	86.6 ± 0.1	86.6 ± 0.1	86.3 ± 0.1
Wearing Necktie	94.9 ± 0.1	95.1 ± 0.1	95.3 ± 0.1	95.5 ± 0.1	95.6 ± 0.1	95.4 ± 0.0	95.2 ± 0.1
Young	87.1 ± 0.2	87.4 ± 0.2	87.7 ± 0.1	88.1 ± 0.2	88.7 ± 0.1	88.5 ± 0.1	88.0 ± 0.2

Table J.3. AUC, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	82.4 ± 0.7	82.9 ± 0.8	83.3 ± 1.0	83.5 ± 0.9	83.7 ± 0.7	83.4 ± 0.6	82.0 ± 0.7
Arched Eyebrows	78.1 ± 0.5	78.9 ± 0.4	79.2 ± 0.7	79.9 ± 0.6	80.4 ± 0.5	80.1 ± 0.4	78.8 ± 0.3
Attractive	80.5 ± 0.2	81.6 ± 0.1	81.8 ± 0.3	82.6 ± 0.1	83.1 ± 0.1	83.1 ± 0.2	82.7 ± 0.1
Bags Under Eyes	72.5 ± 0.5	73.4 ± 0.7	73.8 ± 0.8	74.6 ± 0.5	74.6 ± 0.8	74.4 ± 0.7	72.6 ± 0.5
Bald	85.6 ± 1.3	85.6 ± 1.6	86.8 ± 1.2	87.4 ± 0.6	86.8 ± 0.8	87.3 ± 1.5	83.4 ± 1.1
Bangs	90.8 ± 0.2	91.4 ± 0.2	91.6 ± 0.4	91.8 ± 0.2	91.8 ± 0.2	91.6 ± 0.3	90.9 ± 0.3
Big Lips	60.8 ± 0.4	61.3 ± 0.4	61.2 ± 0.3	61.9 ± 0.2	61.5 ± 0.3	60.1 ± 0.3	58.5 ± 0.5
Big Nose	73.9 ± 0.6	74.2 ± 0.7	74.7 ± 0.6	75.0 ± 0.4	75.3 ± 0.4	74.9 ± 0.6	73.3 ± 0.4
Black Hair	85.1 ± 0.2	85.6 ± 0.2	85.9 ± 0.5	86.4 ± 0.4	86.7 ± 0.5	86.3 ± 0.5	85.7 ± 0.3
Blond Hair	89.8 ± 0.4	90.4 ± 0.3	90.5 ± 0.3	90.8 ± 0.2	90.7 ± 0.3	90.6 ± 0.4	89.9 ± 0.3
Blurry	74.9 ± 0.5	75.4 ± 1.0	75.5 ± 0.7	75.1 ± 0.6	74.6 ± 1.0	73.9 ± 0.7	72.6 ± 1.1
Brown Hair	79.6 ± 0.4	81.1 ± 0.5	81.6 ± 0.4	82.5 ± 0.4	83.0 ± 0.2	82.5 ± 0.1	81.9 ± 0.5
Bushy Eyebrows	80.1 ± 0.9	80.8 ± 1.3	81.0 ± 1.2	81.5 ± 0.7	81.2 ± 0.9	80.8 ± 0.9	80.0 ± 0.7
Chubby	73.7 ± 0.6	74.5 ± 0.3	75.2 ± 0.6	74.7 ± 0.4	74.0 ± 0.7	73.4 ± 0.6	70.9 ± 0.6
Double Chin	71.9 ± 0.4	72.9 ± 0.4	73.7 ± 0.8	73.2 ± 0.5	72.6 ± 0.8	71.6 ± 0.6	69.0 ± 1.0
Eyeglasses	98.0 ± 0.3	98.2 ± 0.2	98.2 ± 0.2	98.3 ± 0.1	98.4 ± 0.2	98.3 ± 0.2	97.8 ± 0.4
Goatee	86.7 ± 0.5	87.6 ± 0.8	87.7 ± 0.8	88.4 ± 0.9	88.9 ± 0.8	88.5 ± 0.9	86.5 ± 1.2
Gray Hair	84.4 ± 0.3	84.5 ± 0.3	85.1 ± 0.6	85.0 ± 0.7	84.7 ± 0.5	85.0 ± 0.7	83.7 ± 1.0
Heavy Makeup	90.7 ± 0.2	90.9 ± 0.1	91.1 ± 0.1	91.6 ± 0.2	91.8 ± 0.1	91.7 ± 0.1	91.5 ± 0.1
High Cheekbones	86.3 ± 0.2	87.0 ± 0.1	87.2 ± 0.1	87.6 ± 0.1	88.0 ± 0.1	87.8 ± 0.1	87.3 ± 0.1
Male	98.3 ± 0.1	98.3 ± 0.1	98.3 ± 0.1	98.3 ± 0.1	98.2 ± 0.1	98.1 ± 0.1	97.6 ± 0.1
Mouth Slightly Open	93.5 ± 0.1	93.7 ± 0.1	93.9 ± 0.1	94.0 ± 0.1	94.2 ± 0.1	94.1 ± 0.1	93.7 ± 0.1
Mustache	72.6 ± 1.6	72.7 ± 0.8	72.6 ± 1.1	72.6 ± 0.9	72.9 ± 0.8	70.4 ± 0.8	69.5 ± 1.6
Narrow Eyes	65.1 ± 0.5	65.1 ± 0.3	65.2 ± 0.2	65.0 ± 0.3	64.4 ± 0.4	63.9 ± 0.4	62.7 ± 0.5
No Beard	90.9 ± 0.6	91.5 ± 0.5	91.6 ± 0.6	91.9 ± 0.4	92.0 ± 0.4	91.9 ± 0.4	91.1 ± 0.6
Oval Face	63.6 ± 0.4	64.6 ± 0.5	64.8 ± 0.5	65.4 ± 0.5	65.4 ± 0.4	64.5 ± 0.5	63.4 ± 0.3
Pale Skin	76.3 ± 0.5	76.5 ± 0.6	77.0 ± 1.0	76.1 ± 0.7	75.7 ± 1.1	74.4 ± 0.8	73.3 ± 0.9
Pointy Nose	66.2 ± 0.3	67.2 ± 0.4	67.7 ± 0.2	68.0 ± 0.4	68.4 ± 0.3	67.8 ± 0.2	66.6 ± 0.2
Receding Hairline	74.4 ± 0.7	74.8 ± 0.8	75.2 ± 0.9	75.2 ± 0.4	75.1 ± 0.5	74.5 ± 0.4	71.8 ± 0.7
Rosy Cheeks	77.8 ± 1.4	78.3 ± 1.6	78.2 ± 1.6	78.7 ± 0.6	78.4 ± 0.8	77.8 ± 1.0	76.5 ± 1.0
Sideburns	86.2 ± 0.8	87.2 ± 1.1	87.3 ± 0.7	87.3 ± 1.1	87.2 ± 1.0	87.1 ± 1.1	86.1 ± 1.1
Smiling	92.3 ± 0.1	92.5 ± 0.1	92.8 ± 0.1	92.9 ± 0.1	93.1 ± 0.1	93.0 ± 0.1	92.8 ± 0.1
Straight Hair	70.3 ± 0.3	71.4 ± 0.4	71.4 ± 0.5	72.0 ± 0.3	71.3 ± 0.2	70.6 ± 0.1	68.1 ± 0.2
Wavy Hair	78.6 ± 0.2	79.8 ± 0.3	80.2 ± 0.4	80.7 ± 0.4	81.0 ± 0.3	80.6 ± 0.3	79.7 ± 0.2
Wearing Earrings	83.8 ± 0.2	84.6 ± 0.6	84.9 ± 0.2	85.2 ± 0.3	85.3 ± 0.4	84.9 ± 0.4	83.4 ± 0.5
Wearing Hat	93.4 ± 0.5	93.5 ± 0.5	93.7 ± 0.3	93.7 ± 0.4	93.8 ± 0.5	93.2 ± 0.4	92.2 ± 0.4
Wearing Lipstick	93.6 ± 0.1	94.0 ± 0.1	94.0 ± 0.1	94.3 ± 0.1	94.4 ± 0.1	94.4 ± 0.2	94.2 ± 0.1
Wearing Necklace	60.0 ± 0.7	60.2 ± 0.5	59.5 ± 0.8	58.5 ± 0.6	56.8 ± 0.7	54.8 ± 0.3	52.2 ± 0.3
Wearing Necktie	76.3 ± 0.4	77.0 ± 0.7	77.9 ± 0.6	78.3 ± 0.2	78.1 ± 0.5	77.5 ± 0.2	74.8 ± 0.8
Young	80.0 ± 0.2	80.5 ± 0.5	80.8 ± 0.5	81.2 ± 0.2	81.6 ± 0.2	81.0 ± 0.3	79.7 ± 0.4

Table J.4. Expected Calibration Error, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	2.9 ± 0.1	2.1 ± 0.1	1.8 ± 0.1	1.2 ± 0.1	0.6 ± 0.1	0.6 ± 0.1	0.7 ± 0.1
Arched Eyebrows	10.2 ± 0.1	7.6 ± 0.1	6.5 ± 0.2	4.8 ± 0.1	2.6 ± 0.1	1.6 ± 0.3	1.6 ± 0.2
Attractive	10.6 ± 0.2	6.8 ± 0.2	5.5 ± 0.4	3.6 ± 0.1	2.0 ± 0.2	1.9 ± 0.2	1.9 ± 0.2
Bags Under Eyes	10.0 ± 0.2	7.4 ± 0.1	6.4 ± 0.1	4.4 ± 0.2	1.9 ± 0.1	1.4 ± 0.2	1.6 ± 0.1
Bald	0.4 ± 0.1	0.3 ± 0.0	0.3 ± 0.0	0.3 ± 0.1	0.3 ± 0.0	0.3 ± 0.1	0.4 ± 0.1
Bangs	2.0 ± 0.1	1.5 ± 0.1	1.1 ± 0.1	0.7 ± 0.1	0.4 ± 0.1	0.5 ± 0.1	0.8 ± 0.1
Big Lips	21.1 ± 0.2	17.5 ± 0.3	16.0 ± 0.3	12.7 ± 0.2	8.6 ± 0.5	7.4 ± 0.6	7.1 ± 0.7
Big Nose	10.9 ± 0.2	9.1 ± 0.3	7.8 ± 0.2	6.1 ± 0.3	3.4 ± 0.3	2.6 ± 0.6	2.5 ± 0.6
Black Hair	5.6 ± 0.1	4.0 ± 0.1	3.4 ± 0.2	2.4 ± 0.1	1.4 ± 0.5	1.4 ± 0.7	1.3 ± 0.5
Blond Hair	1.8 ± 0.0	1.3 ± 0.1	1.0 ± 0.1	0.6 ± 0.0	0.4 ± 0.1	0.4 ± 0.1	0.4 ± 0.1
Blurry	2.1 ± 0.1	1.5 ± 0.1	1.3 ± 0.1	1.0 ± 0.1	0.7 ± 0.1	0.5 ± 0.1	0.5 ± 0.1
Brown Hair	6.4 ± 0.2	4.2 ± 0.3	3.4 ± 0.3	3.1 ± 0.5	3.2 ± 0.4	3.2 ± 0.4	3.6 ± 0.4
Bushy Eyebrows	4.0 ± 0.1	2.8 ± 0.1	2.2 ± 0.2	1.3 ± 0.1	0.8 ± 0.3	0.8 ± 0.4	1.2 ± 0.4
Chubby	2.5 ± 0.1	2.0 ± 0.1	1.8 ± 0.1	1.4 ± 0.1	0.9 ± 0.0	0.8 ± 0.1	0.8 ± 0.2
Double Chin	2.0 ± 0.1	1.7 ± 0.1	1.5 ± 0.1	1.1 ± 0.1	0.6 ± 0.1	0.5 ± 0.1	0.5 ± 0.2
Eyeglasses	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.2 ± 0.0
Goatee	1.1 ± 0.1	0.8 ± 0.1	0.7 ± 0.1	0.8 ± 0.1	0.9 ± 0.2	0.9 ± 0.2	1.1 ± 0.2
Gray Hair	0.8 ± 0.1	0.7 ± 0.0	0.6 ± 0.0	0.4 ± 0.1	0.4 ± 0.1	0.5 ± 0.1	0.7 ± 0.1
Heavy Makeup	4.5 ± 0.2	3.4 ± 0.2	2.9 ± 0.1	2.0 ± 0.1	1.0 ± 0.1	0.7 ± 0.1	0.7 ± 0.1
High Cheekbones	7.3 ± 0.2	5.2 ± 0.2	4.3 ± 0.1	3.1 ± 0.0	1.8 ± 0.1	1.5 ± 0.2	1.7 ± 0.0
Male	0.5 ± 0.1	0.5 ± 0.0	0.5 ± 0.1	0.5 ± 0.1	0.4 ± 0.1	0.3 ± 0.1	0.3 ± 0.1
Mouth Slightly Open	3.0 ± 0.1	2.2 ± 0.1	1.9 ± 0.1	1.4 ± 0.1	0.7 ± 0.1	0.7 ± 0.1	0.7 ± 0.2
Mustache	1.6 ± 0.1	1.4 ± 0.1	1.2 ± 0.1	1.0 ± 0.1	0.6 ± 0.1	0.5 ± 0.1	0.5 ± 0.1
Narrow Eyes	8.8 ± 0.2	7.2 ± 0.1	6.4 ± 0.1	5.1 ± 0.1	3.8 ± 0.2	3.5 ± 0.3	3.1 ± 0.5
No Beard	1.8 ± 0.2	1.4 ± 0.1	1.2 ± 0.1	0.8 ± 0.1	0.5 ± 0.1	0.5 ± 0.1	0.6 ± 0.1
Oval Face	18.2 ± 0.1	13.9 ± 0.4	12.2 ± 0.2	9.4 ± 0.3	4.5 ± 0.1	2.3 ± 0.3	1.2 ± 0.2
Pale Skin	1.5 ± 0.1	1.1 ± 0.1	0.8 ± 0.1	0.7 ± 0.1	0.4 ± 0.1	0.5 ± 0.0	0.5 ± 0.1
Pointy Nose	14.6 ± 0.3	9.6 ± 0.4	7.6 ± 0.1	5.1 ± 0.1	2.5 ± 0.2	2.0 ± 0.1	1.4 ± 0.2
Receding Hairline	3.7 ± 0.1	2.8 ± 0.1	2.4 ± 0.1	1.8 ± 0.1	1.0 ± 0.1	0.6 ± 0.1	0.7 ± 0.1
Rosy Cheeks	2.1 ± 0.0	1.4 ± 0.1	1.1 ± 0.0	0.8 ± 0.1	0.6 ± 0.2	0.6 ± 0.2	0.7 ± 0.2
Sideburns	1.0 ± 0.1	0.8 ± 0.1	0.7 ± 0.0	0.5 ± 0.0	0.3 ± 0.1	0.4 ± 0.2	0.4 ± 0.2
Smiling	3.6 ± 0.1	2.7 ± 0.1	2.2 ± 0.1	1.6 ± 0.2	0.9 ± 0.1	0.7 ± 0.2	0.6 ± 0.1
Straight Hair	10.1 ± 0.1	6.9 ± 0.2	5.7 ± 0.2	3.8 ± 0.1	1.6 ± 0.1	0.9 ± 0.2	0.8 ± 0.0
Wavy Hair	10.2 ± 0.2	7.2 ± 0.4	5.9 ± 0.3	4.6 ± 0.2	3.7 ± 0.4	3.5 ± 0.4	3.3 ± 0.4
Wearing Earrings	5.1 ± 0.1	3.6 ± 0.2	2.9 ± 0.2	2.1 ± 0.0	1.1 ± 0.2	0.9 ± 0.2	1.1 ± 0.1
Wearing Hat	0.4 ± 0.0	0.4 ± 0.0	0.4 ± 0.0	0.3 ± 0.0	0.3 ± 0.0	0.3 ± 0.0	0.3 ± 0.1
Wearing Lipstick	3.0 ± 0.0	2.3 ± 0.1	2.2 ± 0.1	2.1 ± 0.1	2.1 ± 0.1	2.2 ± 0.2	2.1 ± 0.2
Wearing Necklace	9.2 ± 0.2	6.5 ± 0.1	5.3 ± 0.2	3.6 ± 0.1	1.4 ± 0.1	0.8 ± 0.2	0.7 ± 0.3
Wearing Necktie	2.1 ± 0.1	1.6 ± 0.0	1.2 ± 0.0	0.8 ± 0.1	0.4 ± 0.1	0.4 ± 0.2	0.6 ± 0.1
Young	8.2 ± 0.1	6.9 ± 0.1	6.0 ± 0.1	4.5 ± 0.0	2.4 ± 0.2	1.7 ± 0.3	1.8 ± 0.3

Table J.5. Threshold Calibration Bias, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	92.7 ± 4.3	92.1 ± 3.6	91.8 ± 3.8	91.1 ± 3.6	89.4 ± 2.6	88.6 ± 2.4	84.9 ± 3.1
Arched Eyebrows	99.4 ± 3.6	99.5 ± 3.6	99.1 ± 3.9	99.9 ± 3.0	99.0 ± 2.8	98.2 ± 2.9	96.1 ± 2.3
Attractive	97.0 ± 0.9	96.8 ± 1.0	96.3 ± 0.9	96.3 ± 0.5	96.2 ± 0.6	95.8 ± 0.8	94.9 ± 0.6
Bags Under Eyes	92.3 ± 2.8	91.1 ± 3.5	90.6 ± 3.7	89.4 ± 2.5	85.2 ± 3.6	83.2 ± 3.3	77.2 ± 1.3
Bald	93.9 ± 3.5	94.4 ± 5.5	97.8 ± 4.8	98.3 ± 3.9	97.4 ± 4.2	99.2 ± 5.7	90.8 ± 5.4
Bangs	96.7 ± 0.9	96.9 ± 0.9	96.7 ± 1.5	96.4 ± 0.8	95.6 ± 1.0	94.9 ± 1.2	93.7 ± 0.5
Big Lips	62.7 ± 4.2	61.7 ± 3.9	60.1 ± 3.8	57.1 ± 2.0	47.8 ± 2.4	38.3 ± 2.1	31.3 ± 2.7
Big Nose	98.4 ± 4.2	96.6 ± 5.3	95.6 ± 5.3	94.5 ± 3.5	91.0 ± 3.3	88.2 ± 3.4	84.1 ± 4.1
Black Hair	94.8 ± 2.8	94.0 ± 1.8	94.0 ± 2.8	94.2 ± 2.8	93.6 ± 2.7	92.9 ± 2.8	92.6 ± 2.0
Blond Hair	97.1 ± 2.0	96.8 ± 1.5	96.7 ± 1.7	96.6 ± 1.4	95.3 ± 0.9	95.4 ± 1.8	94.7 ± 1.2
Blurry	82.1 ± 3.6	80.4 ± 4.2	80.2 ± 3.7	76.1 ± 3.4	73.2 ± 4.3	70.6 ± 3.0	67.6 ± 3.9
Brown Hair	107.6 ± 2.1	106.6 ± 4.0	105.3 ± 3.1	105.9 ± 2.6	104.0 ± 2.3	102.9 ± 1.9	103.0 ± 2.7
Bushy Eyebrows	90.6 ± 4.6	89.2 ± 5.7	87.8 ± 5.6	86.9 ± 3.6	84.7 ± 4.1	82.4 ± 4.0	80.9 ± 3.8
Chubby	87.7 ± 1.9	86.8 ± 1.7	87.5 ± 2.2	84.9 ± 1.9	80.6 ± 2.1	77.5 ± 2.9	69.8 ± 2.7
Double Chin	77.5 ± 3.2	77.5 ± 3.1	78.0 ± 3.9	75.3 ± 2.2	71.0 ± 4.5	66.7 ± 2.8	60.7 ± 4.0
Eyeglasses	98.1 ± 0.6	98.5 ± 0.7	98.5 ± 0.3	98.9 ± 0.4	98.7 ± 0.3	98.8 ± 0.6	98.0 ± 1.0
Goatee	107.4 ± 2.7	107.4 ± 2.4	107.6 ± 3.3	110.7 ± 3.0	111.8 ± 4.7	112.2 ± 4.1	110.6 ± 4.9
Gray Hair	98.8 ± 2.1	97.0 ± 1.6	97.1 ± 1.7	95.8 ± 2.3	93.1 ± 2.1	94.8 ± 2.3	95.0 ± 3.7
Heavy Makeup	100.5 ± 0.3	100.3 ± 0.6	100.1 ± 0.4	100.7 ± 0.3	100.9 ± 0.1	100.9 ± 0.4	102.3 ± 0.6
High Cheekbones	98.2 ± 1.5	98.1 ± 1.3	97.9 ± 1.3	97.8 ± 1.4	97.9 ± 1.4	97.6 ± 1.1	98.4 ± 0.7
Male	99.3 ± 0.2	99.2 ± 0.3	99.2 ± 0.3	99.2 ± 0.3	99.2 ± 0.4	99.2 ± 0.4	99.2 ± 0.4
Mouth Slightly Open	99.4 ± 0.5	99.4 ± 0.5	99.2 ± 0.4	99.3 ± 0.4	99.5 ± 0.4	99.2 ± 0.3	99.3 ± 0.8
Mustache	73.1 ± 4.4	72.9 ± 2.5	71.4 ± 3.1	69.8 ± 3.4	69.7 ± 4.0	60.2 ± 1.0	57.4 ± 6.9
Narrow Eyes	59.3 ± 2.0	56.3 ± 1.4	55.4 ± 2.0	51.5 ± 1.6	45.8 ± 1.5	42.5 ± 1.7	39.4 ± 2.4
No Beard	95.1 ± 2.1	95.6 ± 1.8	95.7 ± 2.0	95.7 ± 1.4	95.6 ± 1.9	94.8 ± 1.5	93.9 ± 2.9
Oval Face	84.0 ± 4.0	81.1 ± 2.9	78.6 ± 3.9	74.2 ± 2.6	64.3 ± 3.3	56.1 ± 2.9	51.5 ± 2.5
Pale Skin	80.0 ± 4.0	78.5 ± 3.2	77.2 ± 4.7	73.6 ± 2.8	70.0 ± 3.7	66.2 ± 2.6	64.6 ± 4.3
Pointy Nose	84.9 ± 2.9	81.3 ± 3.0	78.8 ± 2.4	74.0 ± 2.2	70.1 ± 1.7	66.1 ± 0.9	63.4 ± 1.6
Receding Hairline	83.3 ± 3.2	81.8 ± 3.4	81.1 ± 4.3	79.3 ± 2.2	76.6 ± 2.3	73.1 ± 2.1	66.6 ± 3.9
Rosy Cheeks	88.9 ± 6.2	88.0 ± 7.0	85.6 ± 7.6	86.2 ± 3.0	83.9 ± 3.6	81.6 ± 5.1	78.9 ± 3.9
Sideburns	94.9 ± 3.4	96.5 ± 3.5	95.3 ± 3.8	95.8 ± 4.2	94.9 ± 4.1	95.6 ± 5.2	95.2 ± 5.2
Smiling	99.8 ± 0.5	99.8 ± 0.8	99.4 ± 0.6	99.3 ± 0.8	98.8 ± 0.8	98.7 ± 0.7	98.3 ± 0.5
Straight Hair	87.2 ± 2.0	85.6 ± 1.1	83.4 ± 1.5	81.1 ± 2.2	75.5 ± 1.8	72.5 ± 0.6	65.6 ± 1.2
Wavy Hair	83.5 ± 0.9	84.0 ± 1.2	83.9 ± 1.3	83.1 ± 1.2	81.9 ± 1.5	81.2 ± 1.5	79.7 ± 1.1
Wearing Earrings	97.0 ± 1.5	97.4 ± 2.0	96.9 ± 1.7	96.5 ± 1.5	95.6 ± 2.0	94.5 ± 1.9	91.2 ± 2.0
Wearing Hat	97.4 ± 1.5	97.4 ± 1.6	97.5 ± 1.0	97.6 ± 1.5	97.1 ± 2.0	96.1 ± 1.4	94.4 ± 1.5
Wearing Lipstick	97.6 ± 0.3	97.9 ± 0.3	97.8 ± 0.4	98.0 ± 0.3	98.3 ± 0.4	98.7 ± 0.3	99.1 ± 0.4
Wearing Necklace	66.0 ± 4.1	58.1 ± 3.9	51.8 ± 3.6	42.4 ± 2.2	28.3 ± 2.8	18.9 ± 1.1	9.3 ± 1.1
Wearing Necktie	82.8 ± 1.9	81.3 ± 2.6	81.9 ± 2.6	81.5 ± 0.6	78.8 ± 2.4	78.3 ± 1.1	71.3 ± 3.2
Young	85.0 ± 1.3	86.3 ± 1.9	85.3 ± 1.9	84.7 ± 0.7	82.3 ± 0.7	80.6 ± 1.4	76.6 ± 1.4

Table J.6. Uncertainty, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	9.3 ± 0.4	10.9 ± 0.4	11.7 ± 0.4	12.9 ± 0.4	14.9 ± 0.4	16.5 ± 0.4	18.5 ± 0.4
Arched Eyebrows	24.3 ± 0.3	30.8 ± 0.4	33.4 ± 0.1	37.5 ± 0.3	43.4 ± 0.5	47.2 ± 0.4	50.5 ± 0.4
Attractive	28.8 ± 0.5	37.6 ± 0.6	40.8 ± 0.4	45.0 ± 0.5	50.0 ± 0.3	52.5 ± 0.3	54.2 ± 0.4
Bags Under Eyes	22.1 ± 0.4	28.1 ± 0.3	30.6 ± 0.2	35.0 ± 0.4	40.9 ± 0.3	44.8 ± 0.3	47.2 ± 0.5
Bald	2.1 ± 0.1	2.2 ± 0.1	2.2 ± 0.1	2.5 ± 0.1	2.7 ± 0.1	2.9 ± 0.1	3.1 ± 0.2
Bangs	7.5 ± 0.2	8.5 ± 0.2	9.0 ± 0.1	10.0 ± 0.1	11.0 ± 0.2	11.8 ± 0.1	12.7 ± 0.3
Big Lips	29.5 ± 1.0	39.7 ± 1.2	44.0 ± 1.4	51.8 ± 1.3	63.2 ± 1.3	69.8 ± 1.1	74.8 ± 1.2
Big Nose	21.1 ± 0.3	26.7 ± 0.7	29.5 ± 0.9	34.9 ± 0.7	42.4 ± 0.6	46.9 ± 0.8	49.9 ± 0.8
Black Hair	17.7 ± 0.4	21.3 ± 0.3	22.7 ± 0.3	24.9 ± 0.3	28.3 ± 0.4	30.2 ± 0.4	32.7 ± 0.4
Blond Hair	8.0 ± 0.1	8.9 ± 0.1	9.4 ± 0.2	10.2 ± 0.2	11.2 ± 0.2	12.0 ± 0.2	12.8 ± 0.2
Blurry	6.1 ± 0.3	7.3 ± 0.5	7.7 ± 0.4	8.2 ± 0.3	9.2 ± 0.4	9.8 ± 0.3	10.4 ± 0.4
Brown Hair	22.5 ± 0.3	28.5 ± 0.5	30.9 ± 0.4	34.1 ± 0.6	37.8 ± 0.5	39.1 ± 0.6	41.0 ± 0.5
Bushy Eyebrows	13.2 ± 0.3	16.0 ± 0.4	17.1 ± 0.5	18.7 ± 0.4	20.9 ± 0.5	22.3 ± 0.8	24.2 ± 1.0
Chubby	7.3 ± 0.3	8.3 ± 0.2	8.8 ± 0.3	9.7 ± 0.2	11.2 ± 0.2	12.4 ± 0.3	13.4 ± 0.3
Double Chin	6.3 ± 0.2	7.0 ± 0.1	7.4 ± 0.3	8.2 ± 0.2	9.3 ± 0.2	10.3 ± 0.4	11.0 ± 0.5
Eyeglasses	0.9 ± 0.1	0.7 ± 0.0	0.7 ± 0.0	0.7 ± 0.1	0.6 ± 0.0	0.7 ± 0.1	1.1 ± 0.2
Goatee	5.1 ± 0.1	5.7 ± 0.1	6.0 ± 0.1	6.7 ± 0.2	7.5 ± 0.1	8.1 ± 0.4	9.3 ± 0.3
Gray Hair	3.5 ± 0.1	3.7 ± 0.1	3.8 ± 0.1	4.2 ± 0.1	4.7 ± 0.1	5.2 ± 0.1	5.8 ± 0.2
Heavy Makeup	14.3 ± 0.1	16.9 ± 0.2	18.0 ± 0.3	19.9 ± 0.2	22.5 ± 0.2	24.4 ± 0.3	26.3 ± 0.2
High Cheekbones	20.3 ± 0.3	25.3 ± 0.5	27.3 ± 0.5	30.0 ± 0.4	33.7 ± 0.6	36.3 ± 0.4	38.2 ± 0.5
Male	3.0 ± 0.1	2.9 ± 0.1	3.0 ± 0.1	3.2 ± 0.1	3.5 ± 0.1	4.3 ± 0.1	6.0 ± 0.1
Mouth Slightly Open	10.7 ± 0.3	12.3 ± 0.1	12.9 ± 0.3	13.9 ± 0.2	15.5 ± 0.1	16.6 ± 0.3	18.1 ± 0.6
Mustache	4.8 ± 0.1	5.4 ± 0.2	5.7 ± 0.2	6.3 ± 0.2	7.2 ± 0.2	7.8 ± 0.2	8.5 ± 0.2
Narrow Eyes	14.9 ± 0.4	19.0 ± 0.3	20.8 ± 0.5	23.8 ± 0.5	27.4 ± 0.7	28.6 ± 0.8	31.0 ± 1.3
No Beard	6.9 ± 0.2	7.6 ± 0.3	7.9 ± 0.2	8.6 ± 0.2	9.4 ± 0.2	10.3 ± 0.3	11.4 ± 0.4
Oval Face	34.6 ± 0.9	46.1 ± 1.3	50.8 ± 1.6	58.0 ± 0.9	70.6 ± 1.0	78.8 ± 0.6	85.3 ± 0.4
Pale Skin	4.8 ± 0.2	5.7 ± 0.3	6.0 ± 0.2	6.3 ± 0.3	7.0 ± 0.3	7.4 ± 0.3	7.7 ± 0.4
Pointy Nose	37.1 ± 0.2	49.8 ± 0.9	54.0 ± 0.7	58.8 ± 0.7	64.7 ± 0.6	67.8 ± 0.4	72.4 ± 0.6
Receding Hairline	9.9 ± 0.2	11.6 ± 0.2	12.6 ± 0.3	13.7 ± 0.3	15.7 ± 0.3	16.9 ± 0.3	18.3 ± 0.4
Rosy Cheeks	10.0 ± 0.3	11.5 ± 0.5	12.3 ± 0.5	13.2 ± 0.2	14.4 ± 0.4	15.2 ± 0.3	16.3 ± 0.3
Sideburns	3.8 ± 0.0	4.3 ± 0.2	4.5 ± 0.1	4.9 ± 0.1	5.5 ± 0.2	6.0 ± 0.2	7.0 ± 0.2
Smiling	12.9 ± 0.3	15.2 ± 0.3	16.0 ± 0.2	17.4 ± 0.3	19.3 ± 0.4	20.7 ± 0.3	22.7 ± 0.5
Straight Hair	26.7 ± 0.5	34.1 ± 0.4	37.0 ± 0.5	41.3 ± 0.4	46.9 ± 0.4	50.3 ± 0.5	54.4 ± 0.5
Wavy Hair	25.9 ± 0.3	32.7 ± 0.7	35.4 ± 0.4	39.1 ± 0.4	43.8 ± 0.5	46.8 ± 0.3	50.6 ± 0.3
Wearing Earrings	16.7 ± 0.2	19.8 ± 0.1	21.3 ± 0.3	23.2 ± 0.3	25.9 ± 0.3	27.5 ± 0.5	29.8 ± 0.3
Wearing Hat	1.4 ± 0.0	1.5 ± 0.0	1.5 ± 0.1	1.7 ± 0.0	1.9 ± 0.1	2.1 ± 0.1	2.4 ± 0.2
Wearing Lipstick	12.6 ± 0.2	14.9 ± 0.4	15.8 ± 0.3	16.9 ± 0.2	18.4 ± 0.2	19.4 ± 0.2	20.3 ± 0.4
Wearing Necklace	24.3 ± 0.8	31.5 ± 0.9	34.5 ± 0.6	39.1 ± 0.8	45.4 ± 1.0	49.3 ± 1.0	52.5 ± 0.6
Wearing Necktie	9.6 ± 0.3	10.4 ± 0.1	11.1 ± 0.3	11.8 ± 0.1	12.6 ± 0.4	13.6 ± 0.1	15.8 ± 0.4
Young	14.9 ± 0.3	18.1 ± 0.4	19.9 ± 0.2	23.3 ± 0.4	29.0 ± 0.3	32.5 ± 0.4	35.9 ± 0.5

Table J.7. Interdependence, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	43.7 ± 0.5	44.9 ± 1.1	45.3 ± 0.6	45.9 ± 0.9	47.1 ± 0.5	48.5 ± 0.6	48.8 ± 1.1
Arched Eyebrows	34.5 ± 0.9	35.4 ± 1.1	36.0 ± 1.1	37.3 ± 0.7	38.2 ± 0.5	39.3 ± 0.7	42.1 ± 1.2
Attractive	44.6 ± 0.2	46.8 ± 0.5	48.2 ± 0.2	50.5 ± 0.4	53.1 ± 0.3	54.5 ± 0.4	56.0 ± 0.6
Bags Under Eyes	28.1 ± 1.0	30.3 ± 0.9	30.9 ± 1.1	32.4 ± 0.7	33.6 ± 0.8	35.1 ± 1.1	36.6 ± 0.7
Bald	13.0 ± 0.6	13.0 ± 0.9	13.8 ± 0.5	13.6 ± 0.6	13.3 ± 0.5	13.9 ± 0.6	14.4 ± 0.8
Bangs	10.5 ± 0.4	10.9 ± 0.4	11.4 ± 0.3	11.6 ± 0.2	12.2 ± 0.3	12.5 ± 0.2	12.9 ± 0.2
Big Lips	15.4 ± 0.5	16.6 ± 0.5	17.5 ± 0.2	19.3 ± 0.3	21.7 ± 0.9	23.4 ± 0.6	24.8 ± 0.8
Big Nose	35.7 ± 0.6	37.9 ± 0.9	39.0 ± 0.5	40.6 ± 0.5	44.0 ± 0.2	46.6 ± 0.5	49.8 ± 0.4
Black Hair	29.0 ± 0.7	29.4 ± 0.6	29.7 ± 0.3	30.0 ± 0.3	30.3 ± 0.6	30.5 ± 0.6	30.8 ± 0.5
Blond Hair	23.8 ± 0.5	24.2 ± 0.7	24.3 ± 0.4	25.2 ± 0.4	25.5 ± 0.4	25.5 ± 0.2	25.3 ± 0.4
Blurry	9.2 ± 0.2	9.8 ± 0.4	10.0 ± 0.2	10.2 ± 0.2	10.3 ± 0.3	10.4 ± 0.4	10.4 ± 0.5
Brown Hair	24.8 ± 0.5	25.5 ± 1.0	25.8 ± 0.6	26.7 ± 0.5	26.9 ± 0.2	27.2 ± 0.3	27.7 ± 0.7
Bushy Eyebrows	18.2 ± 0.8	19.0 ± 0.9	19.3 ± 0.6	19.2 ± 0.5	20.0 ± 0.6	20.7 ± 0.5	21.7 ± 0.5
Chubby	40.8 ± 1.6	44.0 ± 1.0	45.1 ± 0.6	47.6 ± 0.6	48.6 ± 1.3	50.6 ± 0.9	52.9 ± 1.2
Double Chin	39.6 ± 1.3	42.7 ± 0.8	43.8 ± 0.8	46.4 ± 0.5	47.3 ± 1.6	49.1 ± 1.2	51.0 ± 2.0
Eyeglasses	14.9 ± 0.2	15.3 ± 0.2	15.3 ± 0.5	15.7 ± 0.4	16.8 ± 0.5	17.2 ± 0.2	17.8 ± 0.3
Goatee	47.3 ± 1.3	48.1 ± 1.2	49.1 ± 1.4	50.4 ± 1.4	52.7 ± 2.0	55.2 ± 1.8	59.8 ± 3.3
Gray Hair	21.2 ± 0.5	21.1 ± 0.5	21.5 ± 0.3	21.7 ± 0.3	22.3 ± 0.8	23.2 ± 0.9	25.3 ± 1.3
Heavy Makeup	69.7 ± 0.2	69.9 ± 0.3	70.2 ± 0.5	70.9 ± 0.2	71.5 ± 0.3	71.8 ± 0.3	73.1 ± 0.3
High Cheekbones	60.4 ± 0.5	62.4 ± 0.6	63.5 ± 0.4	65.2 ± 0.4	67.2 ± 0.3	68.6 ± 0.3	71.8 ± 0.7
Male	74.2 ± 0.6	74.4 ± 0.6	74.5 ± 0.6	74.7 ± 0.4	75.0 ± 0.4	75.4 ± 0.4	75.5 ± 0.2
Mouth Slightly Open	33.4 ± 0.1	33.7 ± 0.4	33.9 ± 0.4	34.2 ± 0.3	34.5 ± 0.3	35.0 ± 0.4	35.1 ± 0.6
Mustache	26.7 ± 1.8	27.0 ± 0.7	27.1 ± 0.9	26.7 ± 1.2	27.4 ± 1.4	26.0 ± 0.8	30.0 ± 3.4
Narrow Eyes	5.6 ± 0.3	6.3 ± 0.3	6.4 ± 0.4	6.5 ± 0.4	6.4 ± 0.3	6.6 ± 0.6	7.0 ± 0.6
No Beard	64.9 ± 0.5	65.8 ± 0.5	65.8 ± 0.4	66.4 ± 0.6	67.1 ± 0.3	67.6 ± 0.6	67.3 ± 0.6
Oval Face	16.3 ± 0.5	18.6 ± 0.6	19.3 ± 0.7	20.4 ± 0.7	21.5 ± 0.5	21.4 ± 0.3	24.7 ± 1.1
Pale Skin	3.7 ± 0.2	3.7 ± 0.3	3.8 ± 0.2	4.0 ± 0.2	4.2 ± 0.2	4.3 ± 0.2	4.5 ± 0.4
Pointy Nose	15.7 ± 0.5	17.4 ± 0.5	18.3 ± 0.5	19.7 ± 0.6	22.1 ± 0.6	23.9 ± 0.5	27.1 ± 0.5
Receding Hairline	17.4 ± 0.8	17.6 ± 0.9	18.4 ± 1.1	19.0 ± 0.8	20.4 ± 0.6	20.8 ± 0.6	22.5 ± 1.2
Rosy Cheeks	18.2 ± 1.1	18.9 ± 1.2	18.8 ± 1.2	19.9 ± 0.4	20.9 ± 0.9	21.7 ± 1.4	24.5 ± 0.7
Sideburns	38.8 ± 1.2	39.8 ± 1.3	40.3 ± 1.6	40.9 ± 1.7	42.6 ± 1.6	45.0 ± 2.3	49.6 ± 2.8
Smiling	63.1 ± 0.5	64.6 ± 0.6	65.6 ± 0.6	67.0 ± 0.6	68.7 ± 0.2	69.8 ± 0.3	72.5 ± 1.0
Straight Hair	17.0 ± 0.6	17.6 ± 0.4	17.6 ± 0.5	18.0 ± 0.4	18.1 ± 0.4	18.5 ± 0.3	17.7 ± 0.8
Wavy Hair	28.4 ± 0.4	29.0 ± 0.5	29.2 ± 0.6	29.4 ± 0.2	29.3 ± 0.6	29.9 ± 0.5	29.6 ± 0.3
Wearing Earrings	25.1 ± 0.7	25.8 ± 0.7	25.8 ± 0.4	26.3 ± 0.5	27.2 ± 0.2	28.0 ± 0.6	28.8 ± 0.4
Wearing Hat	12.2 ± 0.2	12.3 ± 0.2	12.5 ± 0.3	12.7 ± 0.2	12.7 ± 0.2	12.8 ± 0.4	12.6 ± 0.4
Wearing Lipstick	80.1 ± 0.2	80.5 ± 0.4	80.6 ± 0.3	81.0 ± 0.1	81.3 ± 0.1	81.4 ± 0.2	81.8 ± 0.3
Wearing Necklace	12.4 ± 0.6	13.7 ± 0.9	14.2 ± 1.3	14.8 ± 1.1	14.8 ± 1.0	13.3 ± 0.9	9.7 ± 1.3
Wearing Necktie	21.4 ± 0.9	20.8 ± 1.0	21.1 ± 0.8	21.6 ± 0.6	21.8 ± 0.8	23.1 ± 0.5	24.2 ± 0.6
Young	39.2 ± 0.5	40.2 ± 0.4	41.3 ± 0.5	42.9 ± 0.6	45.7 ± 0.3	46.8 ± 0.7	46.4 ± 0.5

Table J.8. 'Male' Bias Amplification, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	–	–	–	–	–	–	–
Arched Eyebrows	2.3 ± 0.4	2.4 ± 0.2	2.4 ± 0.2	2.7 ± 0.3	3.0 ± 0.5	3.8 ± 0.5	5.1 ± 0.2
Attractive	-0.0 ± 0.3	0.1 ± 0.2	0.3 ± 0.2	0.6 ± 0.2	0.8 ± 0.2	1.0 ± 0.3	1.6 ± 0.2
Bags Under Eyes	1.9 ± 0.5	2.4 ± 0.8	3.1 ± 0.5	4.1 ± 0.2	5.4 ± 0.9	6.6 ± 0.6	8.8 ± 0.8
Bald	–	–	–	–	–	–	–
Bangs	-0.1 ± 0.2	-0.0 ± 0.4	0.1 ± 0.4	0.3 ± 0.1	0.1 ± 0.2	0.1 ± 0.2	-0.3 ± 0.2
Big Lips	2.4 ± 0.2	1.7 ± 0.6	1.9 ± 0.5	2.0 ± 0.8	-0.1 ± 0.8	-2.1 ± 1.4	-5.5 ± 2.3
Big Nose	3.3 ± 0.7	3.8 ± 0.7	4.4 ± 0.2	5.4 ± 1.0	7.9 ± 1.2	10.0 ± 0.7	12.1 ± 0.4
Black Hair	0.1 ± 0.5	-0.0 ± 0.4	0.3 ± 0.5	0.2 ± 0.4	0.3 ± 0.4	0.3 ± 0.6	0.2 ± 0.5
Blond Hair	2.4 ± 0.4	2.4 ± 0.3	2.4 ± 0.3	2.8 ± 0.2	3.1 ± 0.3	3.1 ± 0.3	3.6 ± 0.2
Blurry	0.9 ± 0.6	0.3 ± 1.0	0.3 ± 1.0	-0.7 ± 1.3	-0.9 ± 1.5	-1.4 ± 0.7	-1.4 ± 0.9
Brown Hair	0.2 ± 0.4	-0.4 ± 0.3	0.3 ± 0.1	0.3 ± 0.5	0.5 ± 0.3	1.0 ± 0.2	1.1 ± 0.5
Bushy Eyebrows	3.4 ± 0.6	5.0 ± 0.7	5.8 ± 0.7	5.6 ± 0.5	7.3 ± 0.6	8.2 ± 0.9	8.4 ± 1.5
Chubby	4.3 ± 0.6	4.4 ± 0.5	4.9 ± 1.0	5.1 ± 0.8	6.5 ± 0.4	8.0 ± 0.8	10.8 ± 0.1
Double Chin	5.3 ± 0.7	4.8 ± 0.8	5.5 ± 1.3	6.0 ± 1.1	7.2 ± 1.4	8.1 ± 0.9	10.3 ± 0.2
Eyeglasses	0.2 ± 0.1	0.2 ± 0.2	0.1 ± 0.1	-0.0 ± 0.2	0.0 ± 0.1	-0.0 ± 0.2	0.1 ± 0.2
Goatee	–	–	–	–	–	–	–
Gray Hair	3.6 ± 1.0	3.6 ± 1.0	4.1 ± 0.8	3.8 ± 0.5	4.4 ± 0.8	4.9 ± 0.8	6.1 ± 0.5
Heavy Makeup	0.1 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.0	0.2 ± 0.0
High Cheekbones	-0.4 ± 0.5	-0.4 ± 0.4	-0.2 ± 0.3	-0.0 ± 0.4	-0.0 ± 0.4	0.2 ± 0.3	-0.1 ± 0.3
Male	–	–	–	–	–	–	–
Mouth Slightly Open	-0.1 ± 0.2	-0.2 ± 0.1	-0.2 ± 0.0	-0.1 ± 0.1	-0.1 ± 0.1	-0.0 ± 0.0	-0.1 ± 0.1
Mustache	–	–	–	–	–	–	–
Narrow Eyes	–	–	–	–	–	–	–
No Beard	-0.6 ± 0.2	-0.5 ± 0.2	-0.5 ± 0.2	-0.5 ± 0.2	-0.5 ± 0.2	-0.6 ± 0.2	-0.7 ± 0.3
Oval Face	4.2 ± 0.6	4.3 ± 0.7	5.1 ± 0.4	6.3 ± 0.5	9.8 ± 0.7	13.3 ± 0.6	17.0 ± 0.9
Pale Skin	1.5 ± 1.2	1.4 ± 1.1	2.2 ± 1.0	3.4 ± 0.8	4.9 ± 0.7	4.9 ± 0.6	5.3 ± 0.8
Pointy Nose	4.5 ± 0.5	5.8 ± 0.5	6.5 ± 0.6	7.7 ± 0.2	9.4 ± 0.5	11.5 ± 0.4	13.9 ± 0.3
Receding Hairline	3.3 ± 0.4	3.2 ± 0.8	4.1 ± 1.1	5.5 ± 1.0	5.9 ± 1.5	7.0 ± 0.6	10.1 ± 1.4
Rosy Cheeks	–	–	–	–	–	–	–
Sideburns	–	–	–	–	–	–	–
Smiling	-0.2 ± 0.2	-0.3 ± 0.2	-0.2 ± 0.1	-0.1 ± 0.1	-0.1 ± 0.1	-0.2 ± 0.2	0.2 ± 0.1
Straight Hair	1.7 ± 1.0	2.2 ± 0.5	2.3 ± 0.8	3.0 ± 0.4	2.9 ± 0.5	2.5 ± 0.7	1.7 ± 1.2
Wavy Hair	4.7 ± 0.3	4.7 ± 0.1	4.8 ± 0.2	5.1 ± 0.3	5.6 ± 0.1	6.1 ± 0.3	6.6 ± 0.3
Wearing Earrings	1.6 ± 0.3	1.8 ± 0.2	1.7 ± 0.2	2.0 ± 0.2	2.4 ± 0.1	2.6 ± 0.2	3.1 ± 0.1
Wearing Hat	0.3 ± 0.9	0.3 ± 0.7	0.5 ± 0.6	0.5 ± 0.2	0.9 ± 0.8	1.3 ± 0.6	1.9 ± 0.3
Wearing Lipstick	-0.0 ± 0.1	0.1 ± 0.1	0.1 ± 0.0	0.1 ± 0.1	0.1 ± 0.0	0.1 ± 0.1	0.0 ± 0.1
Wearing Necklace	1.7 ± 0.2	2.4 ± 0.2	2.7 ± 0.3	3.3 ± 0.4	3.9 ± 0.2	3.8 ± 0.1	4.3 ± 0.2
Wearing Necktie	–	–	–	–	–	–	–
Young	0.1 ± 0.2	0.1 ± 0.2	0.2 ± 0.2	0.2 ± 0.1	0.2 ± 0.1	0.3 ± 0.2	0.1 ± 0.2

Table J.9. 'Young' Bias Amplification, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	–	–	–	–	–	–	–
Arched Eyebrows	0.1 ± 0.2	0.0 ± 0.2	-0.0 ± 0.3	0.2 ± 0.1	0.2 ± 0.2	0.3 ± 0.1	0.7 ± 0.2
Attractive	0.1 ± 0.1	0.4 ± 0.3	0.6 ± 0.1	0.8 ± 0.2	1.2 ± 0.1	1.1 ± 0.2	0.7 ± 0.1
Bags Under Eyes	1.4 ± 0.6	1.9 ± 0.6	2.1 ± 0.4	3.0 ± 0.8	3.7 ± 0.5	4.7 ± 0.8	6.3 ± 0.7
Bald	-1.3 ± 0.5	-1.3 ± 0.4	-1.6 ± 0.8	-1.7 ± 1.1	-2.3 ± 0.3	-1.7 ± 0.5	-1.3 ± 0.6
Bangs	0.3 ± 0.4	0.3 ± 0.2	0.2 ± 0.2	0.0 ± 0.4	0.0 ± 0.2	0.0 ± 0.3	-0.2 ± 0.3
Big Lips	2.6 ± 0.4	2.1 ± 0.2	2.0 ± 0.3	2.0 ± 0.4	1.6 ± 0.6	1.6 ± 0.6	2.1 ± 0.4
Big Nose	1.9 ± 0.7	2.7 ± 0.8	2.9 ± 0.8	3.5 ± 0.3	5.4 ± 0.5	7.3 ± 0.7	8.6 ± 1.1
Black Hair	-0.6 ± 0.4	-0.5 ± 0.1	-0.7 ± 0.2	-0.4 ± 0.2	-0.4 ± 0.2	-0.3 ± 0.2	0.0 ± 0.3
Blond Hair	1.0 ± 0.3	1.1 ± 0.3	1.0 ± 0.3	1.0 ± 0.2	0.8 ± 0.4	0.6 ± 0.3	0.4 ± 0.2
Blurry	-1.1 ± 0.5	-0.4 ± 0.6	-0.8 ± 0.6	-0.4 ± 1.0	-0.5 ± 1.1	-0.1 ± 0.6	-0.2 ± 0.8
Brown Hair	-0.2 ± 0.2	-0.3 ± 0.3	-0.2 ± 0.3	-0.0 ± 0.3	0.0 ± 0.3	0.2 ± 0.3	0.2 ± 0.6
Bushy Eyebrows	0.5 ± 0.4	0.4 ± 0.4	0.3 ± 0.4	0.2 ± 0.4	0.0 ± 0.3	0.2 ± 0.2	0.6 ± 0.4
Chubby	0.8 ± 1.2	2.6 ± 0.9	2.9 ± 1.0	3.3 ± 1.2	3.8 ± 1.0	4.4 ± 0.6	6.1 ± 1.1
Double Chin	4.0 ± 0.6	4.6 ± 0.9	5.3 ± 1.4	5.5 ± 1.0	5.6 ± 1.2	6.3 ± 0.5	7.8 ± 1.6
Eyeglasses	-0.4 ± 0.3	-0.3 ± 0.2	-0.2 ± 0.3	-0.3 ± 0.1	-0.2 ± 0.2	-0.3 ± 0.1	-0.2 ± 0.1
Goatee	-1.6 ± 0.7	-2.0 ± 0.9	-2.4 ± 0.9	-1.7 ± 0.7	-1.1 ± 1.1	0.7 ± 0.4	1.4 ± 0.9
Gray Hair	1.6 ± 0.6	1.8 ± 0.3	1.8 ± 0.4	1.7 ± 0.5	1.5 ± 0.2	1.2 ± 0.3	0.7 ± 0.5
Heavy Makeup	-0.4 ± 0.1	-0.3 ± 0.2	-0.2 ± 0.1	-0.2 ± 0.1	-0.1 ± 0.1	0.0 ± 0.1	0.1 ± 0.0
High Cheekbones	–	–	–	–	–	–	–
Male	0.4 ± 0.1	0.4 ± 0.1	0.4 ± 0.0	0.4 ± 0.0	0.4 ± 0.1	0.5 ± 0.1	0.5 ± 0.1
Mouth Slightly Open	–	–	–	–	–	–	–
Mustache	1.5 ± 1.6	1.2 ± 0.5	2.2 ± 1.7	4.2 ± 1.1	4.6 ± 1.5	7.0 ± 1.4	10.0 ± 3.9
Narrow Eyes	0.8 ± 0.8	1.6 ± 0.7	2.0 ± 0.6	2.4 ± 0.8	1.8 ± 1.0	1.7 ± 1.0	2.0 ± 1.1
No Beard	-0.2 ± 0.0	-0.2 ± 0.0	-0.2 ± 0.1	-0.2 ± 0.0	-0.2 ± 0.0	-0.2 ± 0.0	-0.2 ± 0.0
Oval Face	1.1 ± 0.6	1.6 ± 0.7	1.7 ± 0.7	2.9 ± 0.5	5.4 ± 0.7	8.6 ± 0.5	10.0 ± 0.8
Pale Skin	2.0 ± 0.7	1.7 ± 0.6	2.0 ± 0.7	2.0 ± 0.4	3.1 ± 0.5	3.7 ± 0.4	4.0 ± 0.4
Pointy Nose	2.1 ± 0.3	2.7 ± 0.9	2.7 ± 0.6	3.3 ± 0.6	4.0 ± 0.6	4.5 ± 0.5	4.7 ± 0.3
Receding Hairline	3.4 ± 1.1	3.3 ± 1.3	4.9 ± 0.5	5.0 ± 1.1	5.9 ± 1.2	7.0 ± 0.6	10.5 ± 0.3
Rosy Cheeks	-0.0 ± 0.5	-0.1 ± 0.4	0.3 ± 0.4	0.0 ± 0.8	0.2 ± 0.9	0.1 ± 0.9	-0.3 ± 0.6
Sideburns	-2.5 ± 0.6	-1.8 ± 1.1	-1.9 ± 1.1	-1.8 ± 1.2	-1.7 ± 0.7	-2.1 ± 0.8	-2.3 ± 0.7
Smiling	–	–	–	–	–	–	–
Straight Hair	1.4 ± 0.4	1.7 ± 0.4	2.1 ± 0.7	2.5 ± 0.3	2.6 ± 0.5	3.7 ± 0.2	4.8 ± 0.5
Wavy Hair	-0.1 ± 0.1	-0.3 ± 0.2	-0.3 ± 0.1	-0.2 ± 0.2	0.1 ± 0.2	0.2 ± 0.2	0.7 ± 0.2
Wearing Earrings	0.2 ± 0.3	0.1 ± 0.1	-0.0 ± 0.2	0.0 ± 0.4	0.1 ± 0.5	-0.1 ± 0.3	0.3 ± 0.1
Wearing Hat	-0.1 ± 0.5	-0.1 ± 0.3	-0.2 ± 0.2	-0.0 ± 0.4	-0.3 ± 0.2	-0.6 ± 0.2	-1.1 ± 0.4
Wearing Lipstick	-0.1 ± 0.1	-0.1 ± 0.1	-0.1 ± 0.1	-0.1 ± 0.1	-0.1 ± 0.1	0.0 ± 0.1	0.1 ± 0.1
Wearing Necklace	-2.2 ± 0.9	-5.1 ± 1.1	-8.0 ± 0.9	-11.2 ± 1.1	-21.3 ± 1.8	-30.5 ± 2.4	-34.4 ± 3.3
Wearing Necktie	4.6 ± 1.6	3.8 ± 0.7	4.3 ± 1.0	4.1 ± 0.2	4.8 ± 1.1	5.6 ± 0.6	8.7 ± 1.3
Young	–	–	–	–	–	–	–

Table J.10. 'Chubby' Bias Amplification, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	1.1 ± 0.3	1.0 ± 0.3	1.3 ± 0.4	1.5 ± 0.2	1.6 ± 0.2	1.6 ± 0.2	1.7 ± 0.3
Arched Eyebrows	0.4 ± 0.0	0.4 ± 0.1	0.4 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.4 ± 0.1
Attractive	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0	-0.0 ± 0.1	-0.0 ± 0.0	-0.0 ± 0.0	-0.1 ± 0.0
Bags Under Eyes	-0.1 ± 0.2	-0.1 ± 0.3	0.1 ± 0.3	0.5 ± 0.2	1.2 ± 0.4	1.8 ± 0.2	2.4 ± 0.1
Bald	-2.0 ± 1.2	-2.0 ± 1.8	-2.3 ± 0.8	-2.1 ± 0.4	-3.2 ± 0.8	-3.1 ± 0.1	-2.4 ± 0.9
Bangs	-0.2 ± 0.1	-0.2 ± 0.1	-0.1 ± 0.1	-0.1 ± 0.0	-0.1 ± 0.1	-0.1 ± 0.1	-0.1 ± 0.1
Big Lips	–	–	–	–	–	–	–
Big Nose	0.7 ± 0.6	0.9 ± 0.8	1.2 ± 0.7	1.6 ± 0.6	2.8 ± 0.6	4.1 ± 0.6	5.0 ± 0.8
Black Hair	–	–	–	–	–	–	–
Blond Hair	0.3 ± 0.1	0.3 ± 0.1	0.3 ± 0.0	0.4 ± 0.0	0.4 ± 0.1	0.4 ± 0.1	0.4 ± 0.0
Blurry	1.2 ± 0.7	1.0 ± 0.3	1.1 ± 0.3	1.4 ± 0.4	1.4 ± 0.6	1.3 ± 0.3	1.0 ± 0.5
Brown Hair	0.0 ± 0.1	0.0 ± 0.1	0.1 ± 0.1	0.1 ± 0.1	0.2 ± 0.1	0.2 ± 0.0	0.0 ± 0.1
Bushy Eyebrows	–	–	–	–	–	–	–
Chubby	–	–	–	–	–	–	–
Double Chin	-4.1 ± 1.7	-2.3 ± 2.5	-1.1 ± 1.2	0.6 ± 0.9	2.3 ± 1.9	3.3 ± 1.5	1.8 ± 0.7
Eyeglasses	-0.2 ± 0.2	-0.2 ± 0.1	-0.2 ± 0.1	-0.2 ± 0.1	-0.2 ± 0.1	-0.2 ± 0.1	-0.1 ± 0.1
Goatee	-0.5 ± 0.6	-0.3 ± 0.2	-0.8 ± 0.4	-0.2 ± 0.8	-0.2 ± 0.4	0.7 ± 0.5	2.0 ± 0.4
Gray Hair	-2.8 ± 0.7	-2.7 ± 0.7	-2.3 ± 0.7	-1.9 ± 0.6	-1.7 ± 0.4	-1.1 ± 0.5	-0.0 ± 0.8
Heavy Makeup	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0
High Cheekbones	0.2 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.3 ± 0.0	0.3 ± 0.1	0.4 ± 0.1
Male	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
Mouth Slightly Open	0.1 ± 0.1	0.1 ± 0.1	0.1 ± 0.1	0.0 ± 0.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.1
Mustache	-3.7 ± 0.8	-3.5 ± 0.2	-2.7 ± 1.6	-2.9 ± 0.8	-1.1 ± 0.9	2.6 ± 1.6	6.7 ± 2.5
Narrow Eyes	-0.0 ± 0.4	0.4 ± 0.4	0.6 ± 0.4	1.1 ± 0.2	1.1 ± 0.4	1.7 ± 0.6	1.9 ± 0.4
No Beard	-0.0 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0	-0.0 ± 0.0	-0.1 ± 0.0
Oval Face	-0.8 ± 0.2	-1.0 ± 0.1	-1.2 ± 0.3	-1.0 ± 0.2	-0.8 ± 0.4	-0.1 ± 0.2	0.3 ± 0.2
Pale Skin	0.4 ± 0.3	0.3 ± 0.3	0.4 ± 0.1	0.3 ± 0.3	0.9 ± 0.3	1.2 ± 0.2	1.1 ± 0.3
Pointy Nose	0.3 ± 0.2	0.4 ± 0.1	0.5 ± 0.1	0.6 ± 0.0	0.7 ± 0.0	0.8 ± 0.1	0.9 ± 0.0
Receding Hairline	0.5 ± 0.5	0.6 ± 0.9	1.0 ± 0.6	1.4 ± 0.4	2.1 ± 0.7	2.6 ± 0.7	4.5 ± 0.8
Rosy Cheeks	0.8 ± 0.2	0.8 ± 0.3	0.8 ± 0.2	0.9 ± 0.2	0.8 ± 0.1	0.5 ± 0.1	0.4 ± 0.2
Sideburns	-1.1 ± 0.4	-0.8 ± 0.3	-0.7 ± 0.6	-1.0 ± 0.6	-0.8 ± 0.1	-0.6 ± 0.5	-0.3 ± 0.6
Smiling	0.2 ± 0.1	0.2 ± 0.0	0.2 ± 0.0	0.1 ± 0.0	0.2 ± 0.0	0.2 ± 0.1	0.2 ± 0.1
Straight Hair	0.6 ± 0.2	0.6 ± 0.1	0.7 ± 0.3	0.8 ± 0.2	1.0 ± 0.2	1.1 ± 0.3	1.4 ± 0.2
Wavy Hair	0.4 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.2 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.4 ± 0.1
Wearing Earrings	0.1 ± 0.2	0.1 ± 0.2	0.1 ± 0.2	0.1 ± 0.1	0.1 ± 0.0	0.2 ± 0.1	0.4 ± 0.0
Wearing Hat	-0.1 ± 0.3	0.0 ± 0.2	0.1 ± 0.2	0.2 ± 0.3	0.1 ± 0.1	-0.0 ± 0.2	-0.1 ± 0.3
Wearing Lipstick	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	0.0 ± 0.0
Wearing Necklace	0.0 ± 0.2	-0.3 ± 0.2	-0.5 ± 0.3	-0.4 ± 0.4	-1.0 ± 0.5	-1.4 ± 0.3	-3.5 ± 3.7
Wearing Necktie	1.5 ± 0.2	1.7 ± 0.7	2.0 ± 0.7	1.8 ± 0.7	1.5 ± 0.6	2.8 ± 0.4	4.2 ± 0.5
Young	-0.5 ± 0.1	-0.5 ± 0.0	-0.4 ± 0.0	-0.4 ± 0.0	-0.3 ± 0.0	-0.3 ± 0.1	-0.3 ± 0.1

Table J.11. 'Pale Skin' Bias Amplification, Joint CelebA training, ResNet18

Sparsity Attribute	0	80	90	95	98	99	99.5
5 o Clock Shadow	0.0 ± 0.1	-0.0 ± 0.1	-0.0 ± 0.1	0.0 ± 0.2	0.0 ± 0.1	-0.0 ± 0.1	-0.0 ± 0.1
Arched Eyebrows	-1.0 ± 0.2	-1.1 ± 0.2	-0.9 ± 0.1	-0.9 ± 0.1	-0.8 ± 0.1	-0.9 ± 0.1	-1.0 ± 0.2
Attractive	0.1 ± 0.1	0.1 ± 0.1	0.1 ± 0.1	0.2 ± 0.1	0.3 ± 0.1	0.4 ± 0.0	0.4 ± 0.1
Bags Under Eyes	0.5 ± 0.1	0.5 ± 0.1	0.5 ± 0.2	0.6 ± 0.1	0.5 ± 0.1	0.7 ± 0.1	0.9 ± 0.1
Bald	–	–	–	–	–	–	–
Bangs	-0.1 ± 0.1	-0.1 ± 0.2	-0.1 ± 0.1	-0.1 ± 0.1	-0.0 ± 0.1	-0.1 ± 0.2	-0.0 ± 0.1
Big Lips	-0.7 ± 0.2	-0.8 ± 0.2	-0.7 ± 0.2	-0.6 ± 0.3	-0.9 ± 0.3	-1.9 ± 0.3	-2.3 ± 0.1
Big Nose	0.6 ± 0.1	0.8 ± 0.2	0.8 ± 0.1	0.8 ± 0.1	1.1 ± 0.1	1.3 ± 0.0	1.6 ± 0.1
Black Hair	0.0 ± 0.1	-0.2 ± 0.2	-0.0 ± 0.2	-0.0 ± 0.1	0.1 ± 0.1	0.0 ± 0.1	0.1 ± 0.1
Blond Hair	0.1 ± 0.1	-0.1 ± 0.1	0.0 ± 0.1	0.2 ± 0.2	0.2 ± 0.1	0.4 ± 0.2	0.2 ± 0.1
Blurry	0.1 ± 0.1	0.2 ± 0.1	0.2 ± 0.2	0.1 ± 0.2	0.1 ± 0.2	0.3 ± 0.2	0.1 ± 0.2
Brown Hair	-0.9 ± 0.2	-0.8 ± 0.2	-0.7 ± 0.1	-0.7 ± 0.1	-0.7 ± 0.1	-0.7 ± 0.2	-0.9 ± 0.2
Bushy Eyebrows	-0.1 ± 0.2	0.1 ± 0.2	0.2 ± 0.2	0.2 ± 0.1	0.4 ± 0.1	0.4 ± 0.1	0.5 ± 0.1
Chubby	0.2 ± 0.1	0.1 ± 0.4	0.1 ± 0.2	0.1 ± 0.2	0.2 ± 0.1	0.3 ± 0.1	0.6 ± 0.1
Double Chin	0.5 ± 0.2	0.3 ± 0.3	0.4 ± 0.3	0.5 ± 0.2	0.8 ± 0.1	0.7 ± 0.1	0.7 ± 0.2
Eyeglasses	-0.0 ± 0.1	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0
Goatee	–	–	–	–	–	–	–
Gray Hair	0.1 ± 0.4	0.1 ± 0.2	-0.1 ± 0.2	-0.1 ± 0.2	-0.4 ± 0.3	-0.4 ± 0.1	-0.4 ± 0.1
Heavy Makeup	0.5 ± 0.0	0.5 ± 0.1	0.5 ± 0.1	0.5 ± 0.1	0.5 ± 0.1	0.6 ± 0.1	0.6 ± 0.0
High Cheekbones	-0.0 ± 0.1	0.0 ± 0.1	0.0 ± 0.0	0.1 ± 0.0	0.0 ± 0.1	0.0 ± 0.1	-0.0 ± 0.0
Male	0.1 ± 0.1	0.1 ± 0.1	0.1 ± 0.0	0.1 ± 0.1	0.1 ± 0.0	0.1 ± 0.0	0.2 ± 0.0
Mouth Slightly Open	-0.1 ± 0.1	-0.1 ± 0.0	-0.2 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.1	-0.1 ± 0.0	-0.1 ± 0.0
Mustache	–	–	–	–	–	–	–
Narrow Eyes	–	–	–	–	–	–	–
No Beard	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0
Oval Face	-1.3 ± 0.1	-1.4 ± 0.1	-1.4 ± 0.1	-1.4 ± 0.2	-1.4 ± 0.2	-1.7 ± 0.3	-1.4 ± 0.5
Pale Skin	–	–	–	–	–	–	–
Pointy Nose	–	–	–	–	–	–	–
Receding Hairline	0.8 ± 0.1	0.7 ± 0.2	0.9 ± 0.3	1.0 ± 0.2	1.0 ± 0.2	1.2 ± 0.2	1.3 ± 0.1
Rosy Cheeks	–	–	–	–	–	–	–
Sideburns	–	–	–	–	–	–	–
Smiling	-0.2 ± 0.0	-0.2 ± 0.0	-0.2 ± 0.0	-0.2 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.0
Straight Hair	-0.1 ± 0.2	0.0 ± 0.2	0.1 ± 0.2	-0.2 ± 0.3	-0.4 ± 0.3	-0.1 ± 0.3	0.1 ± 0.3
Wavy Hair	-0.1 ± 0.1	-0.1 ± 0.1	-0.2 ± 0.2	-0.1 ± 0.1	-0.1 ± 0.1	-0.1 ± 0.1	-0.1 ± 0.1
Wearing Earrings	0.3 ± 0.1	0.3 ± 0.1	0.3 ± 0.2	0.3 ± 0.1	0.3 ± 0.1	0.4 ± 0.1	0.4 ± 0.2
Wearing Hat	0.8 ± 0.3	0.8 ± 0.2	1.0 ± 0.2	0.9 ± 0.2	1.0 ± 0.3	1.0 ± 0.1	1.2 ± 0.1
Wearing Lipstick	-0.1 ± 0.0	-0.1 ± 0.0	-0.1 ± 0.1	-0.1 ± 0.0	-0.1 ± 0.0	-0.0 ± 0.0	0.0 ± 0.1
Wearing Necklace	–	–	–	–	–	–	–
Wearing Necktie	0.3 ± 0.3	0.3 ± 0.2	0.4 ± 0.2	0.2 ± 0.2	0.2 ± 0.2	0.3 ± 0.1	0.4 ± 0.2
Young	-0.0 ± 0.0	-0.1 ± 0.1	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.0 ± 0.0	-0.1 ± 0.0

K. Results on the Animals with Attributes Dataset

In our efforts of investigating the exacerbation of bias in sparse models, we further validate our results on CelebA on the Animals with Attributes (AwA2) [51] dataset, which consists of 37 322 images of animals belonging to 50 different classes. Each class is annotated using 85 binary attributes, which indicate the presence or absence of different characteristics in each species. We note that AwA2 is not as suited for the study of bias as CelebA, for two important reasons: first, there is a reduced sociological incentive of studying bias, compared to a dataset consisting of human subjects; furthermore, the attributes are labelled at species level, rather than individually per sample, which makes it more difficult to disambiguate between different sources of bias. Nonetheless, we believe AwA2 still serves as a useful validation for our findings on CelebA.

In our experiments with AwA2, we train dense and GMP-RI models at $\{80\%, 90\%, 95\%, 98\%, 99\%, 99.5\%\}$ sparsities to predict the 85 binary attributes. For both the dense and sparse models we use the same training setup and hyperparameters as for CelebA. We follow the original dataset split [51], where the train and test set classes are disjoint: 40 classes are used for training and validation, and the remaining 10 we leave for testing. We follow a different split for train and validation, compared to [51]; namely, we randomly select 80% of the samples for training and the remaining 20% for validation. Our choice is motivated by the fact that further splitting the classes between train and validation would make it more likely to exclude certain attributes from the train set; this would be detrimental to our analysis, as we want to measure the presence of bias on certain attributes. The categories under which it is most sensible to study Categorical bias are not well-established for Animals with Attributes; here we use Furry, Bipedal, Domestic, and Water, where the last refers to the animal's natural habitat.

Our results are shown in Figure K.21. We observe a degradation in AUC scores for models at $\geq 98\%$sparsity, whereas the accuracy does not decrease significantly even at 99.5% sparsity. Moreover, the fraction of uncertain samples increases substantially at $\geq 98\%$ sparsity, and roughly doubles compared to the dense model at 99.5% sparsity. Other metrics, such as TCB or interdependence, decrease slightly with sparsity, compared to the dense model; however, in the case of Systematic (and, to a large extent, Categorical) bias, the fact that the attributes are labeled at the species level - and therefore the model need only learn the species to also learn all the labels - makes this result difficult to interpret. We further study the amplification of bias with sparsity, by following a similar approach to the one on CelebA: namely, we select four category identity attributes with respect to which we compute bias amplification on the remaining attributes. On all attributes considered we did not observe a significant increase in bias induced by sparsity. Generally, our observations on AwA2 seem to validate our findings from CelebA: good quality models even at high sparsity, and substantially increased uncertainty with sparsity.

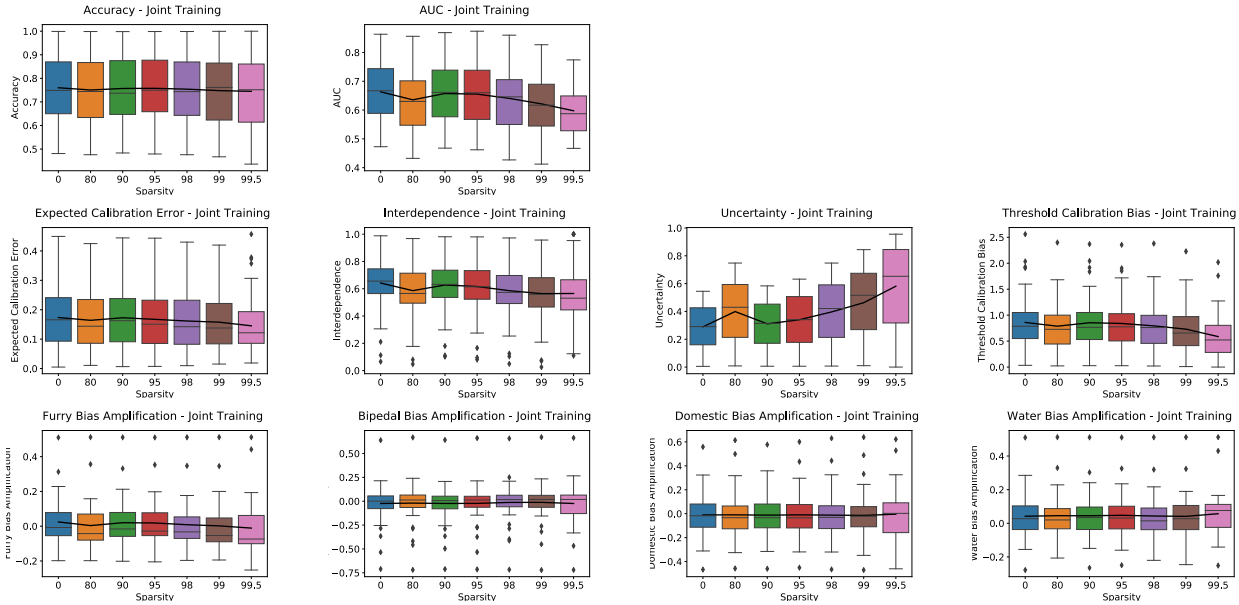

Figure K.21. [Animals With Attributes2 / ResNet18 / GMP-RI] Accuracy and Systematic Bias metrics (TCB, ECE, Interdependence) of ResNet18 models jointly trained on all AwA2 attributes. The thick black line denotes the mean value at each sparsity level.

Metric	Dense	Sparsity (%)			
		80	90	95	98
ID F1 Score (%)	50.1±0.3	51.6±0.6	50.1±1.8	48.2±1.6	43.7±1.2
OOD F1 Score (%)	38.5±1.3	39.8± 0.7	38.9 ± 1.9	37.2 ± 1.6	33.4 ±1.3
ID Precision (%)	54.1±0.8	55.3±0.4	54.6±2.2	52.7±2.5	49.3±1.9
OOD Precision (%)	41.5±0.8	43.2±0.4	43.4 ± 2.2	41.5 ±2.5	38.2 ±1.9
ID Recall (%)	53.1±0.6	53.3±0.7	51.2±1.7	50.4±2.9	45.4±5.6
OOD Recall (%)	39.6±0.7	40.1± 0.7	40.0 ± 1.6	38.6 ± 1.3	35.5 ±1.7

Table L.12. Average ID and OOD Test Accuracy and for iWildcam models

L. iWildcam Results

The iWildCam dataset [3] is a set of images collected from wildlife-spotting camera traps provided by the Wildlife Conservation Society (WCS). Each image contains at least one animal, and is annotated with a single animal label (there is an extension of this dataset containing unlabelled images, but we do not use it here). In total, the dataset contains 203 029 labelled images, divided between a training set, in-distribution (ID) validation and test sets, and out-of-distribution(OOD) validation and test sets. The train (129 809 images), ID validation (7 134 images), and ID test (8154 images) sets were obtained by splitting the photographs from 243 cameras, while the OOD validation (14 961 images) and test (42 791 images) sets were obtained using images from an additional 32 and 48 cameras, respectively. The iWildCam dataset contains images of 182 different animals and is highly unbalanced in terms of class sizes, with some classes having less than 10 images in the training data, and some over 1000. For this reason, the dataset is frequently used to study rare-subgroup performance, as in [3].

We study compression-induced bias on the iWildcam dataset by measuring the performance degradation for rarer classes. It is postulated in, e.g, [29] that features that distinguish rare examples may be cannibalized by larger classes, leading to degraded performance for those classes. To conduct our study, we trained models at 0%, 80%, 90%, 95%, and 98% sparsity. All models used the training settings and hyperparameters (including data augmentations, batch size, epoch number, optimizer, and learning rates) used in [3] for plain ERM. The pruning was done using the GMP-RI variant of Global Magnitude Pruning, with pruning beginning at epoch 2 and ending at epoch 11, with another 2 epochs afterwards for fine-tuning. We use the metrics of Macro Precision, Recall, and F1-Score used in [3]; these metrics assign equal weight to each class when computing the aggregate values. Additionally, we measure the softmax entropy across classes of the predictions as a measure of uncertainty. Ths measure is computed by first computing the softmax per-class prediction for each example,

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}},$$

where the sum is taken over all classes. As these values sum up to 1 for each example, they may be loosely interpreted as the probabilities for each class; thus, their entropy

$$H(X) = -\sum_i \sigma(x_i) \log(\sigma(x_i))$$

may be interpreted as a measure of uncertainty as to the correct class (where the sum is once again taken over that example's predictions for every class). To stay ideologically consistent with the Macro metrics used to evaluate accuracy, we compute the average entropy across examples by upweighting rare class examples, so that each class has equal weight in determining the average entropy.

We report our accuracy and bias results in Table L.12. Following convention, we report Precision, Recall, and F1-score in %, even though F1-score is a hyperbolic mean of the first two. We observe that the Macro F1-Score, precision, and recall stay fairly constant between Dense, 80%, and 90% sparse models, but then decay fairly rapidly after that, with a ID F1-Score drop of 6.4% between 90% sparse and 98% sparse models =, and an OOD F1-Score drop of 5.4%. We also note that precision and recall are fairly well balanced in the models. The dense results are a fairly close match to the results obtained in [3]; we attribute the difference primarily to the choice of random seed.

We additionally break down the dense and sparse F1-Score, Precision, and Recall by the size of the class in the test data, as shown in Figure L.22. We observe that class size has a very large impact on all three metrics, with very small classes having extremely low performance as compared to larger classes. We further observe that, outside of the very low-performant 0-5 class size, sparsity disproportionately affects the performance of smaller classes, with F1-Score decreasing substantially with

sparsity for classes containing 6-50 examples, but remaining nearly constant for classes of over 50 elements on ID test data. On OOD data, the performance decreases with sparsity on all class sizes (again, over 5 examples), but the decrease is greater on smaller class sizes. This experiments provides further evidence for the hypothesis outlined in [29] that ERM with sparsity can sacrifice smaller group performance to preserve accuracy on larger groups. However, we note that on the ID test data, we do not see this effect until the higher sparsity levels of 95% and 98%, where overall F1 score also starts to drop.

The entropy of the models is shown in Figure L.23. We observe that the entropy of the models increases with sparsity when measured on the OOD test set; on the ID test set, the entropy also increases, but only for high-sparsity models where the accuracy is also lower, and the smaller classes' performance is largely decayed. This adds confirmatory evidence that increased uncertainty is related to increased bias as sparsity increases.

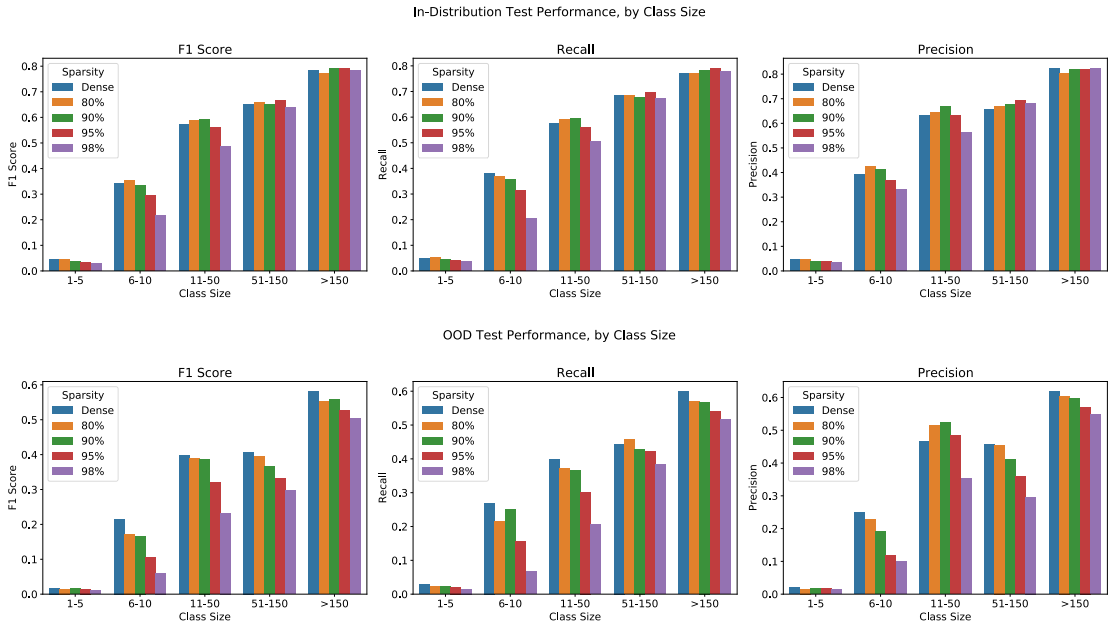

Figure L.22. [iWildCam / ResNet18 / GMP-RI] Macro F1-Score, Precision, and Recall by sparsity and size of test class.

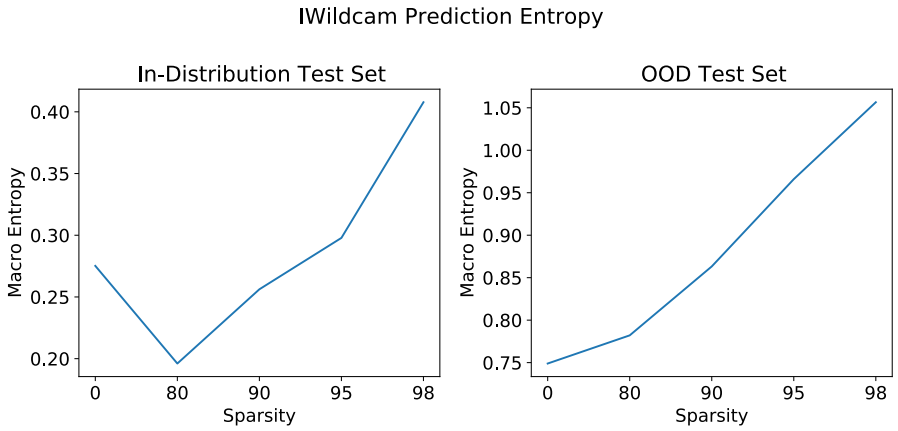

Figure L.23. [iWildCam / ResNet18 / GMP-RI] Average prediction Entropy across sparsities.

M. Example Viewer

As part of our contributions, we provide a simple UI tool that allows the people working with a dataset, for example engineers or scientists who build models, to quickly and easily examine a small subset of the data. This tool is not meant to be a replacement for external review, such as example relabeling, or an audit of the data collection pipeline; if these tools are available than we strongly recommend they be used; however, they can be expensive and difficult to implement; our Example Viewer can serve as a minimum check in case that more extensive review is impossible. Further, our tool relies primarily on random sampling to choose examples to examine. This may cause users of the tool to miss small effects in the data, which may be surfaced by tools using more sophisticated error metrics to choose examples. We also note that other tools already exist that allow for model and dataset exploration, for instance the Kaggle dataset viewer, or HuggingFace Hub. However, unlike these tools, the Example Viewer runs locally. This design choice confers the advantage that neither data nor models need be uploaded to a third-party tool; in addition to increased privacy, this means that it is very easy to integrate the Example Viewer into a research pipeline, where tens or even hundreds of types models may be created as part of the study, and any of them may be instantly auditable through the tool. Finally, the tool is web-based using the popular Flask framework, and so can be run on a development machine (e.g., a laptop), on a development server while still allow for local viewing, or on a world-open server as a regular website. We provide the tool as code, which requires only Python and a few additional packages to run. It is available at [will be made available upon acceptance].

The tool has two core functionalities: viewing a random sample of positive and negative examples for a binary prediction task, and viewing a random selection of true positives, false negatives, false positives, and true negatives for a binary prediction task. These are further stratified by high and low certainty examples, using the definition in section 2.4. In all cases, reloading the page produces a new random sample.

Despite its simplicity, a quick examination can yield clues to defects in the dataset. As case studies, we first present the viewer showing positive and negative examples for the four CelebA identity categories - Male (Figure M.24), Young (Figure M.25), Chubby (Figure M.26), and Pale Skin (Figure M.27). Then, we show three case studies that demonstrate problems in the dataset that can easily be detected from the Example Viewer. Please note that in all illustrations, we avoid cherry-picking by taking the screenshot of the very first returned random set. First, we demonstrate that the categories "Wearing Necklace" (Figures M.32, M.33) and "Wearing Necktie" (Figures M.30, M.31) often cannot be inferred from the cropped version of the CelebA dataset, due to the fact that images are generally cropped at the neck, between the chin and the clavicle. The cropping frequently removes or largely reduces direct visual evidence of the presence or absence of the attribute, leaving the model to use other, correlated features, even though the human raters had access to the full version of the image. Additionally, we show a view of positive and negative examples of the Wearing Lipstick attribute (Figures M.28, M.29). These examples readily show that in many cases it is very difficult to determine whether the person in the photograph is wearing lipstick by only examining the mouth. Rather, it appears far more likely that the human raters used other information in the photograph, such as the gender, clothes, and other makeup of the subject as additional information in choosing the correct label. relying heavily on this information can naturally lead to bias in the human labels, thus making any bias (and accuracy) measurement of the predictions unreliable. A closer examination of the viewer output that also shows correct and incorrect high and low-certainty predictions of the GMP-RI 80% sparse model on these attributes (Figures M.33, M.31, and M.29) confirms this observation. Additionally, we note that in the case of Wearing Lipstick and Wearing Necklace, the high-certainty True Negatives appear to skew much more heavily Male than do the low-certainty True Negatives, and the opposite is true for Wearing Necktie. This suggests that the Male attribute and markers of this attribute are used heavily by the model in order to make these predictions.

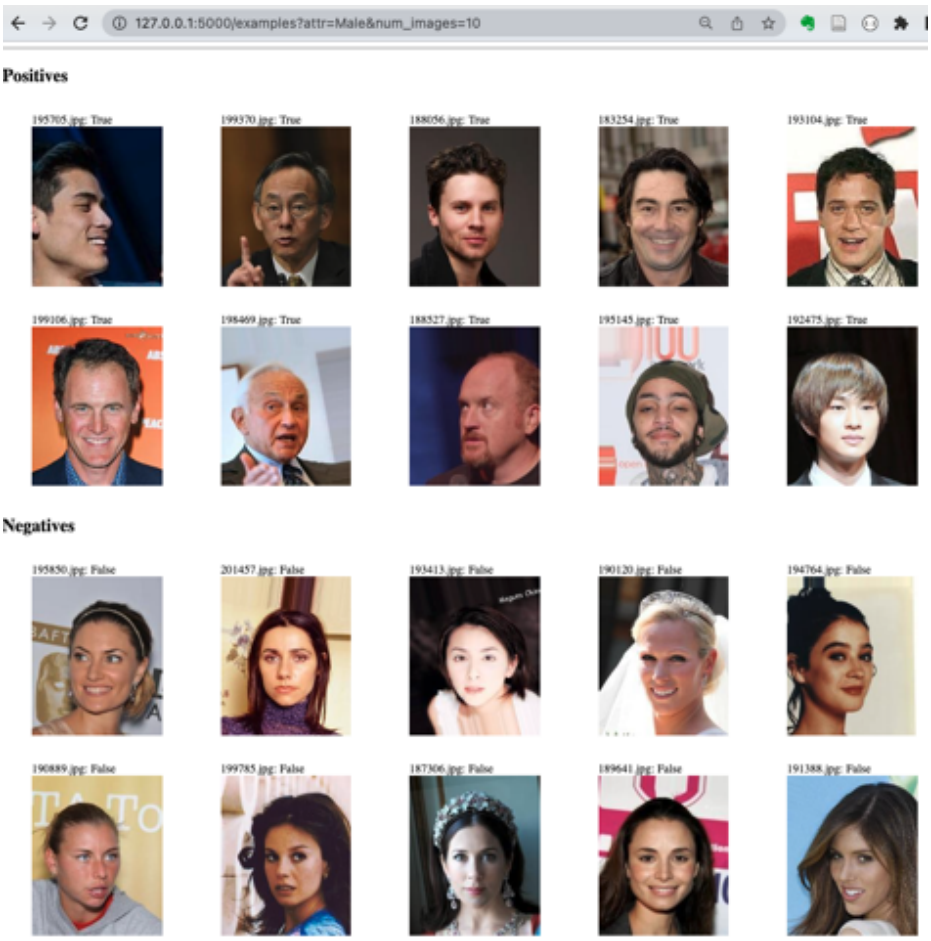

Figure M.24. Examples of images that are Positive and Negative for Male.

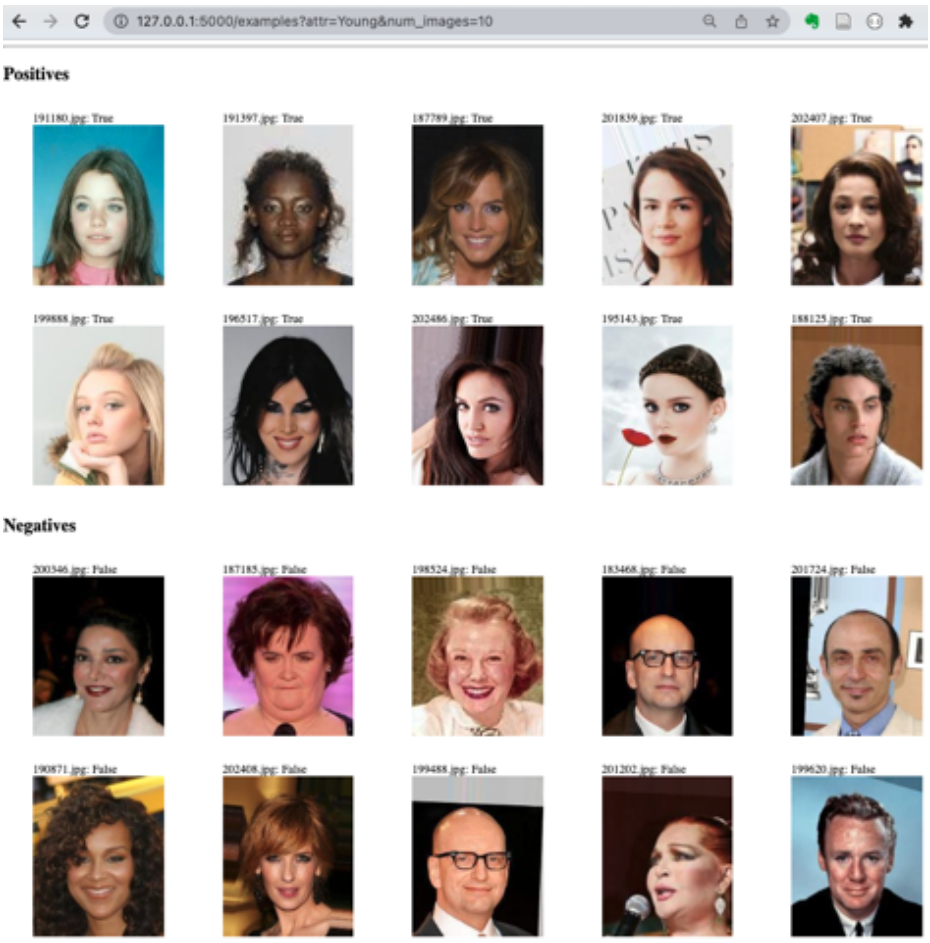

Figure M.25. Examples of images that are Positive and Negative for Young.

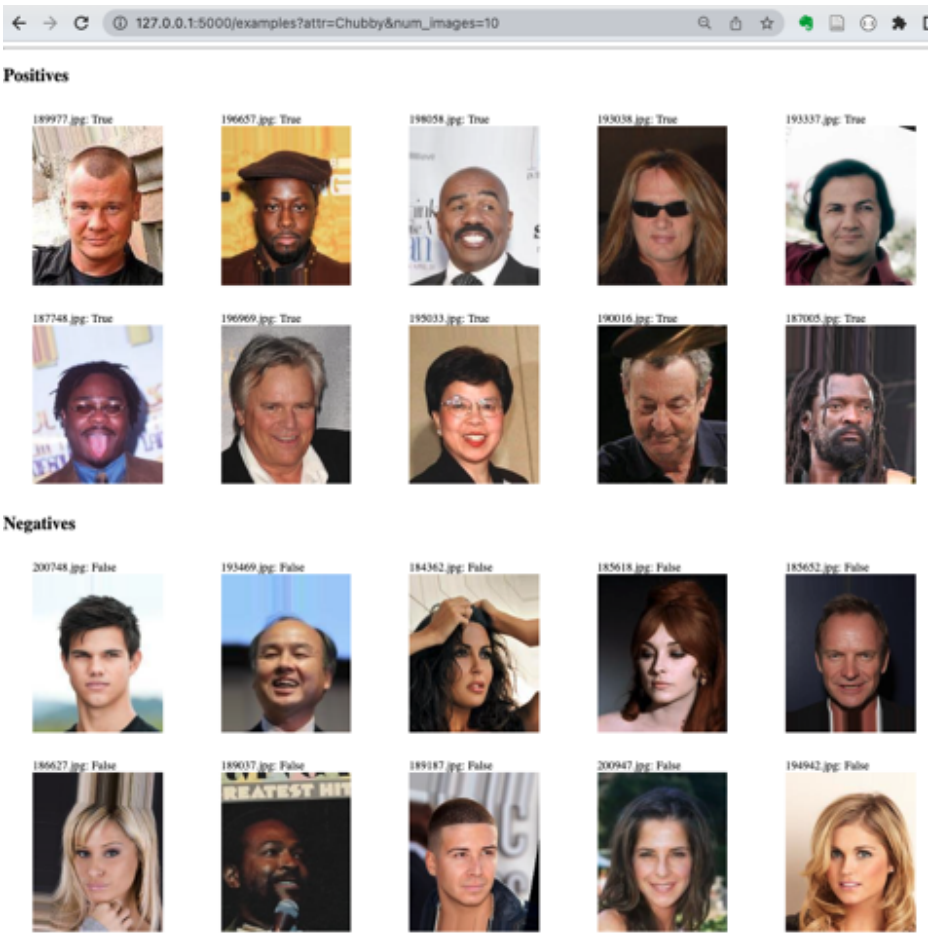

Figure M.26. Examples of images that are Positive and Negative for Chubby.

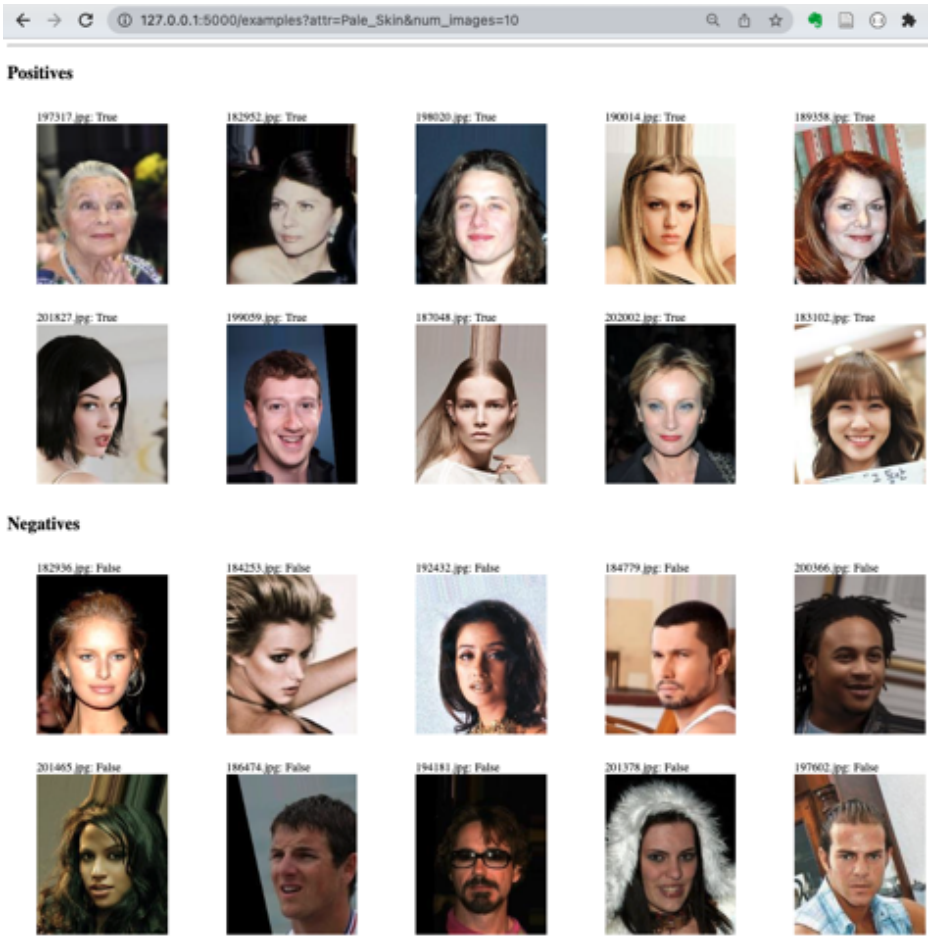

Figure M.27. Examples of images that are Positive and Negative for Pale Skin.

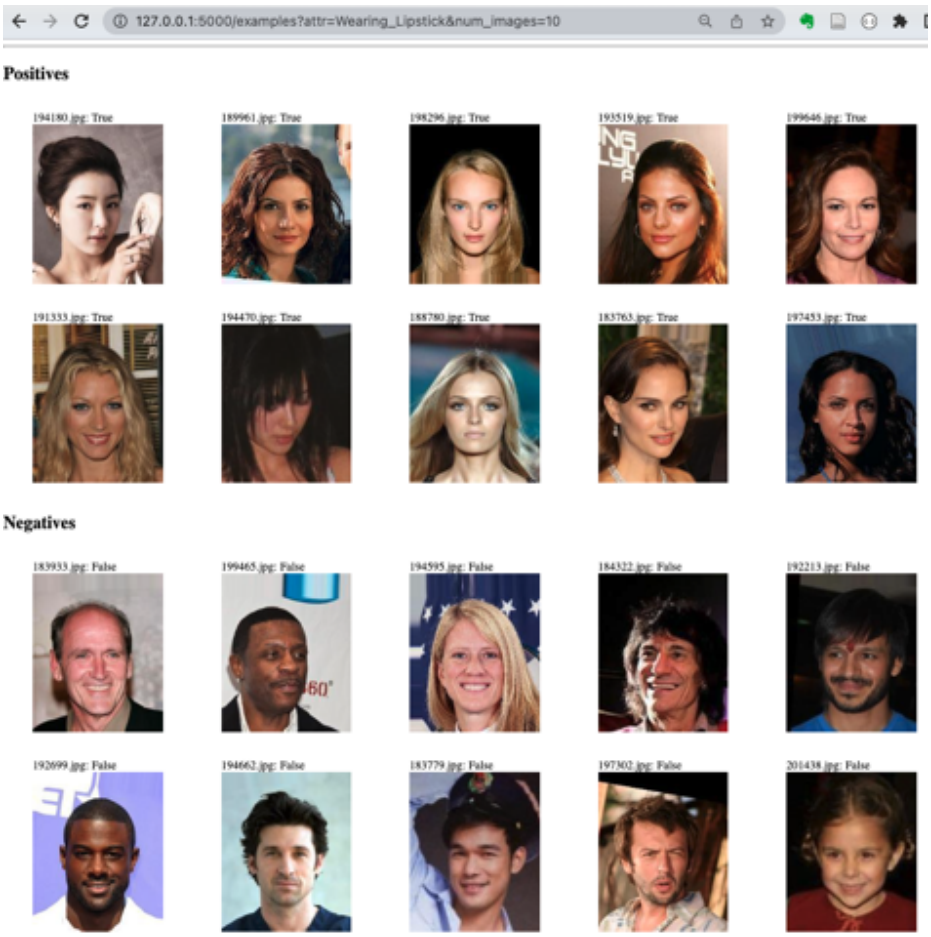

Figure M.28. Examples of images that are Positive and Negative for Wearing Lipstick.

Figure M.29. Examples of 80% sparse model performance on images that are Positive and Negative for Wearing Lipstick.

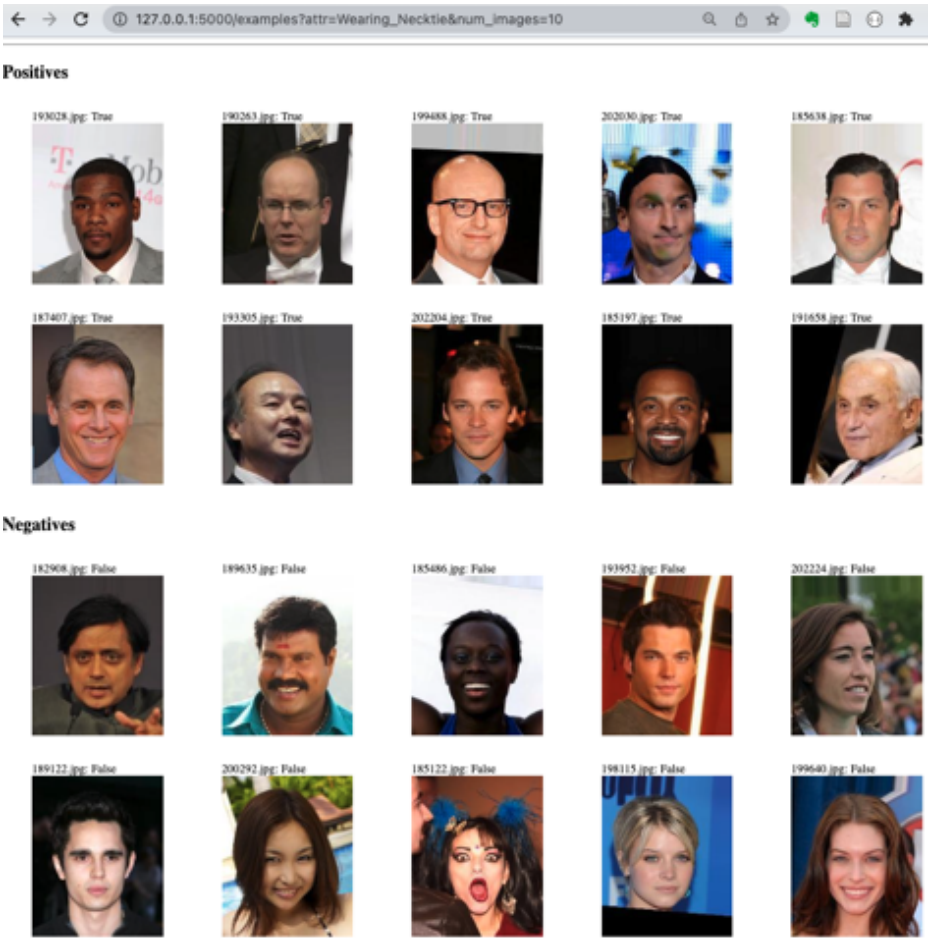

Figure M.30. Examples of images that are Positive and Negative for Wearing Necktie.

Figure M.31. Examples of 80% sparse model performance on images that are Positive and Negative for Wearing Necktie.

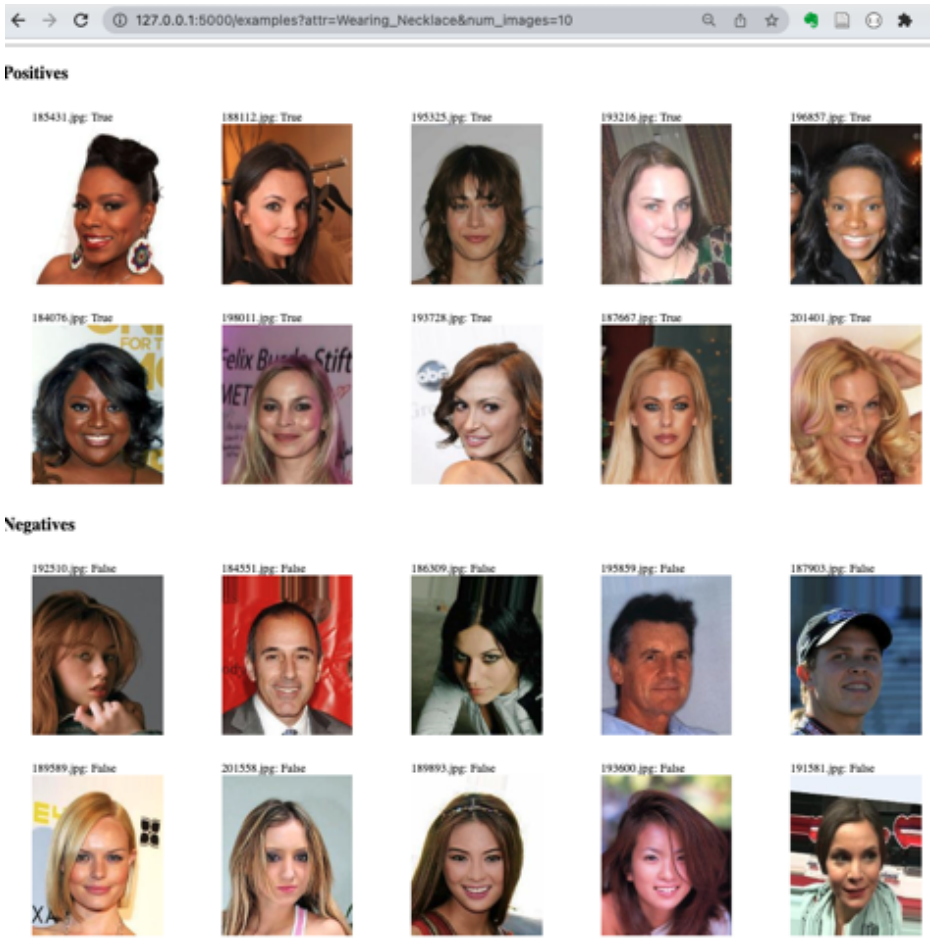

Figure M.32. Examples of images that are Positive and Negative for Wearing Necklace.

Figure M.33. Examples of 80% sparse model performance on images that are Positive and Negative for Wearing Necklace.