Gone With the Bits: Benchmarking Bias in Facial Phenotype Degradation Under Low-Rate Neural Compression

Tian Qiu [*1] **Arjun Nichani** [*2] **Rasta Tadayontahmasebi** [2] **Haewon Jeong** [1]

Abstract

In this study, we investigate how facial phenotypes are distorted under neural image compression and the disparity of this distortion across racial groups. Neural compression methods are gaining popularity due to their impressive rate-distortion performance and their ability to compress to extremely small bitrates, below 0.1 bits per pixel (bpp). As deep learning architectures, these models are prone to bias during the training process, leading to unfair outcomes for individuals in different groups. We first demonstrate, by benchmarking five popular neural compression algorithms, that compressing facial images to low bit-rate regimes leads to the degradation of specific phenotypes (e.g. skin type). Next, we highlight the bias in this phenotype degradation across different race groups. We then show that leveraging a racially balanced dataset does not help mitigate this bias. Finally, we examine the relationship between bias and realism of reconstructed images at different bitrates.

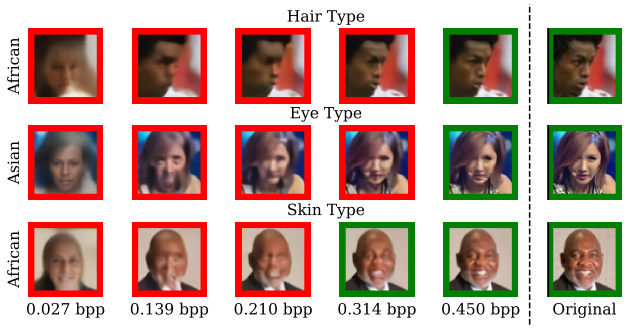

Figure 1: Reconstructed images by the *GaussianMix-Attn* model at different compression rates, trained on CelebA, reveal that the phenotype classifiers consistently misclassify images of African and Asian individuals when predicting hair type and eye type across compression rates. Additionally, the African group suffers from a shift in skin color at lower rates, which is captured by a skin type classifier model.

1. Introduction

Lossy image compression aims to faithfully represent images with a small number of bits. It has been under extensive research for the past 40 years, and image encoders/decoders ("codecs") such as JPEG (Wallace, 1991) have been a crucial enabling technology in the modern digital world. Despite the widespread adoption in daily use, traditional codecs suffer greatly at extreme scenarios with low-bandwidth availability, such as space (Gao et al., 2023), underwater (Li et al., 2023), low-power communication systems (Ez-Zazi et al., 2018) and low-latency systems (Hu & Chen, 2021). These extreme scenarios impose a very narrow information bottleneck that limits the reconstruction quality of traditional codecs. In recent years, neural network-based compression ("neural compression") has enabled image compression under extremely low bitrate scenarios, which is desirable under these scenarios with extreme information bottlenecks.

Regardless of the compression method used, image reconstructions at low bitrates suffer from significant distortion due to the insufficient number of bits passing through the information bottleneck. In this low-bitrate regime, JPEG compression has been shown to exhibit biased performance in facial recognition tasks across different racial and phenotype groups (Yucer et al., 2022a). Neural networks also demonstrate bias in similar settings. When face images are downsampled or corrupted with high levels of noise and then reconstructed by a neural network, the resulting images tend to exhibit distortion in a specific direction: African American faces tend to be reconstructed to appear Caucasian while Caucasian faces maintain their facial features, a phenomenon known as the "White Obama" problem (Jalal et al., 2021; Laszkiewicz et al., 2024). While downsampling or adding noise is not exactly compression, it follows the same essential principle that images pass through a narrow

[*]Equal contribution [1]Department of Electrical and Computer Engineering, [2]Department of Computer Science, University of California, Santa Barbara. Correspondence to: Tian Qiu <tian_qiu@ucsb.edu>, Haewon Jeong <haewon@ece.ucsb.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

information bottleneck and are then reconstructed.

In this paper, we ask: *When using a full neural compression model, consisting of a neural encoder and decoder, would it show similar bias, or would it exhibit even worse bias as it learns from a potentially imbalanced dataset?* Despite extensive research on fairness in machine learning models, to the best of our knowledge, our work is the first to examine fairness in machine learning-based compression models. Unlike existing work (Jalal et al., 2021; Laszkiewicz et al., 2024; Tanjim et al., 2022) that performs a fixed operation to create an information bottleneck (e.g., downsampling) and then applies a neural decoder for reconstruction, our work investigates the bias that could arise in a jointly-trained neural encoder-decoder pair.

In this work, we provide a comprehensive empirical analysis on facial reconstruction bias in state-of-the-art neural compression models. Instead of using conventional distortion measures (e.g., mean square error), we use a facial phenotype classifier to quantify distortion, as we believe this method better captures how an image from one racial group transitions to a different racial group at lower bitrates. Using this framework, we make several interesting observations:

- We benchmark five popular neural compression architectures and demonstrate that, across all, *skin type*, *eye type*, and *hair type* degrade faster than other phenotypes when reducing bitrates.
- We show that for extremely low bitrates, bias can be amplified, particularly for the African race group.
- We demonstrate that, in general, training with race-balanced datasets does not help remove bias in phenotype degradation at extremely low rates.
- We highlight the relationship between bias and realism for different neural compression models and illustrate the sporadic trend at lower bitrates.

2. Related Work

Neural Image Compression In the recent years, we have witnessed rapid advancements in neural compression models. Yang et al. (2023) and (Chen et al., 2024) give a holistic overview of recent works. State-of-the-art neural compression models demonstrate superior rate-distortion performance compared to popular traditional image codecs such as JPEG, BPG (Bellard, 2014), and even the latest hand-engineered codec in VVC (Bross et al., 2021). Early work in this field (Toderici et al., 2015; 2017) utilized recurrent neural networks, while many subsequent studies have employed VAE-based architectures (Townsend et al., 2019; Duan et al., 2023a;b).

Fairness in Face Analysis The processing of facial images is utilized across various domains, including facial biometrics, and facial expression recognition. Fairness in such systems is crucial and has been studied in various aspects of face and biometric analysis (Drozdowski et al., 2020; Vangara et al., 2019; Serna et al., 2019). Buolamwini & Gebru (2018) evaluated commercial gender classification tools and identified that darker-skinned females suffer from significantly higher misclassification rate than lighter-skinned males. Klare et al. (2012) found that various face recognition systems exhibited the poorest performance on cohorts comprising females, Black individuals, and those aged 18-30. Motivated by the imbalanced distribution of datasets used for facial expression detection, Xu et al. (2020) investigate biases across gender, race, and age groups, and propose methods to mitigate these biases in such models.

Fairness in Image Denoising and Upsampling Stemming from the "White Obama" problem, fairness has been explored across image upsampling, denoising, and super-resolution models. Jalal et al. (2021) design fairness definitions and highlight tradeoffs for these types of models. Tanjim et al. (2022) examine the disappearance of minority attributes during image-to-image generation. Laszkiewicz et al. (2024) study the fairness in face image upsampling, demonstrating bias when imbalanced datasets are used while training these upsampling methods.

Fairness in Image Compression Our work is closely related to Yucer et al. (2022a), which studies the impact of JPEG compression on facial verification and identification tasks and the amount of adverse impact of JPEG compression to different racial and phenotype-based subgroups. They define bias as the different amount of downstream task performance degradation across groups. They find phenotype groups of darker skin tones, wide nose, curly hair, and monolid eye shape suffer the most adverse impact in the facial recognition tasks.

3. Problem Definition and Experimental Setup

3.1. Problem Definition

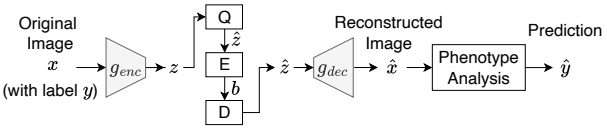

Figure 2: Diagram of the evaluation methodology. Q represents quantization, and E and D represent standard entropy encoders and decoders, b is the compressed bitstream.

Let $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^{n}$ be our dataset, where $x_i \in \mathcal{X}$ is our image, $y_i \in \mathcal{Y}$ is our phenotype label, and $a_i \in \mathcal{A}$ is our protected attribute (race). Our goal is to examine how y_i is preserved in reconstructions of x_i after neural compression

and how this trend differs across $\mathcal{A}$. An overview of our evaluation pipeline can be found in Figure 2.

Neural Compression Neural compression models consist of an encoder $g_{\text{enc}} : \mathcal{X} \to \mathcal{Z}$ and a decoder $g_{\text{dec}} : \mathcal{Z} \to \mathcal{X}$, each built from learnable network layers. For each x_i, the encoder is used to obtain the latent space output z_i, which is then quantized to $\hat{z}_i$ and compressed losslessly to a bitstream b. b is then decompressed to $\hat{z}_i$ and passed through the decoder to provide the reconstruction $\hat{x}_i$.

Phenotype Degradation To examine phenotype degradation in neural compression reconstructions, we train a phenotype classifier. Labelling the phenotypes in these reconstructions manually may be the most accurate way to examine degradation, but a neural classifier acts as a close proxy. This is similar to approaches in previous literature (Jalal et al., 2021; Tanjim et al., 2022).

First, given a dataset $\mathcal{D}$, we split into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ and use $\mathcal{D}_{\text{train}}$ to train a classifier $f : \mathcal{X} \to \mathcal{Y}$ to predict the phenotype labels. Then, given a pretrained encoder and decoder at bitrate r, the original test dataset $\mathcal{D}_{\text{test}}$ is compressed to the bitrate r and reconstructed to $\widehat{\mathcal{D}}_r^{\text{test}}(g_{\text{enc}}, g_{\text{dec}}) = \{(\hat{x}_i, y_i, a_i)\}_{i=1}^n$. To measure phenotype degradation at the given rate, we evaluate the accuracy of f on $\widehat{\mathcal{D}}_r^{\text{test}}$:

$$\text{Acc}(g_{\text{enc}}, g_{\text{dec}}, r) = \mathbb{P}_{(\hat{x},y)\sim\widehat{\mathcal{D}}_r^{\text{test}}(g_{\text{enc}},g_{\text{dec}})}(f(\hat{x}) = y). \quad (1)$$

We further define groupwise accuracy as follows:

$$\text{Acc}(g_{\text{enc}}, g_{\text{dec}}, r|a) = \mathbb{P}_{(\hat{x},y,a)\sim\widehat{\mathcal{D}}_r^{\text{test}}(g_{\text{enc}},g_{\text{dec}})}(f(\hat{x}) = y|A = a). \quad (2)$$

Training details for the phenotype classifier are highlighted in Appendix A.1.

Bias To quantify bias, we leverage *accuracy disparity*, the maximum difference of accuracy across all groups. Given a rate r, an encoder g_{enc}, and a decoder g_{dec}, the bias metric is defined as:

$$\text{Bias}(r, g_{\text{enc}}, g_{\text{dec}}) \triangleq \max_{a,b\in\mathcal{A}} [\text{Acc}(r, g_{\text{enc}}, g_{\text{dec}}|a) - \text{Acc}(r, g_{\text{enc}}, g_{\text{dec}}|b)]. \quad (3)$$

This definition of bias is derived from a popular fairness metric, *accuracy parity*, in which equal accuracies across all groups imply fairness in a classifier (Berk et al., 2017; Zafar et al., 2017).

Remark We acknowledge that these phenotype classifiers can be biased themselves. Following the bias definition in Equation 3, the bias of the classifier at a single rate is captured in $\text{Bias}(r, g_{\text{enc}}, g_{\text{dec}})$, however, our focus of the paper is evaluating how $\text{Bias}(r, g_{\text{enc}}, g_{\text{dec}})$ amplifies as we reduce the rate r. By using a single classifier trained on the clean image data, we can compare bias values across various bitrates (e.g. $\text{Bias}(r_{\text{low}}, g_{\text{enc}}, g_{\text{dec}}) - \text{Bias}(r_{\text{high}}, g_{\text{enc}}, g_{\text{dec}})$) to examine how bias is amplified as bitrates are reduced. This provides insight into the bias induced by the neural compression algorithm. The bias in the classifier is worthy of thorough investigation in the future. We highlight future directions in Section 5.

3.2. Experimental Setup

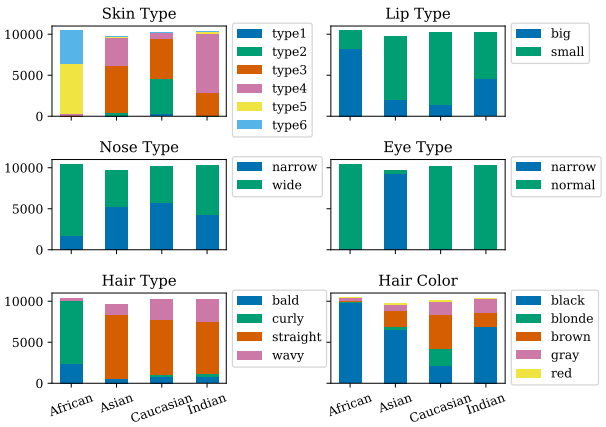

Figure 3: Distribution of phenotype classes for each category across racial groups in RFW dataset.

Datasets For phenotype analysis, we use the Racial Faces in the Wild (RFW) dataset (Wang et al., 2019) and a recently released facial phenotype annotation dataset specifically for RFW (Yucer et al., 2022b). This annotation dataset provides labels for six phenotype categories—eye type, hair type, hair color, lip type, nose type, and skin type—across four racial groups: African, Indian, Asian, and Caucasian. The distribution of phenotypes across these racial groups is depicted in Figure 3.

When measuring bias, we utilize the racial groups as our sensitive attribute, defining $\mathcal{A}$ as the set of all racial groups. When performing inference for multiclass classification tasks (hair color, skin type, and hair type), we group the three most dominant classes for each group. This allows us to evaluate the extent to which phenotypes flipped to those not prevalent in the racial group of the initial image.

We train neural image compression models on both racially balanced and imbalanced datasets. To train these models, we use the CelebA dataset (Liu et al., 2018), which has a significantly imbalanced racial composition with more than 70% of the images from the white racial group (Kärkkäinen & Joo, 2019). Additionally, we leverage the FairFace dataset (Kärkkäinen & Joo, 2019) to investigate the effect of a balanced training dataset. The FairFace dataset contains over

100,000 images with a balanced racial composition across seven race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. Finally, to quantify the relationship between realism and bias, we utilize the DemogPairs dataset (Hupont & Fernández, 2019) as a reference to compute FID scores of decoded images.

Neural Image Compression We evaluate a diverse collection of neural image compression models with different bitrates. We evaluate three fixed-rate models, *Hyperprior* (Ballé et al., 2018), *Joint* (Minnen et al., 2018), and *GaussianMix-Attn* (Cheng et al., 2020). All of these models are trained towards a fixed trade-off between rate and distortion. We train these models to five rates with different operational bitrates. The model proposed in the *QRes* paper (Duan et al., 2023b) is a progressive decoding model that supports encoding images to 12 bitrates with one trained model. This is achieved by encoding only a subset from all the available latent variables. We follow this approach and encode images to 5 different bitrates with progressive decoding. The *VarQRes* model (Duan et al., 2023a) is a variable rate compression model. The network is trained to encode and decode images that lies in a range of rate-distortion trade-off points. We train neural compression models on the CelebA and FairFace datasets. These datasets are chosen to do comparisons between the impact of racially balanced and imbalanced training sets. For the fixed rate models, we adopt the implementations from the CompressAI (Bégaint et al., 2020) library. For the other two models, we adopt the official implementation provided by the authors (Duan et al., 2023b). Training details are listed in Appendix A.2.

4. Evaluation

This section is organized as follows. In Section 4.1 we demonstrate the performance of our phenotype classifier and identify phenotypes lost during low bitrate neural compression. Section 4.2 introduces the racial bias of the neural compression algorithms at low bitrates. Section 4.3 provides a qualitative analysis of image reconstructions across different bitrates. Section 4.4 examines the effect of using a racially balanced dataset. Finally, Section 4.5 discusses bias-realism relationships across the different neural compression models.

4.1. Phenotype Degradation in Neural Compression

First, we examine how phenotype classification accuracies on decoded images change as we reduce the compression bitrate. We first observe significant accuracy decreases across the *skin type*, *eye type*, *hair type* and *lip type* phenotypes as we utilize smaller bitrates. Additionally, phenotypes such as *nose type* and *hair color* experience moderate accuracy decreases. The only exception to this decreasing trend is *skin type*, where the accuracy occasionally increases at the lowest bitrate. This phenomenon occurs sporadically but appears interesting and requires further investigation in future works. The trends we observe are relatively consistent across all of the neural compression architectures we evaluate (Appendix B).

4.2. Phenotype Degradation by Racial Group

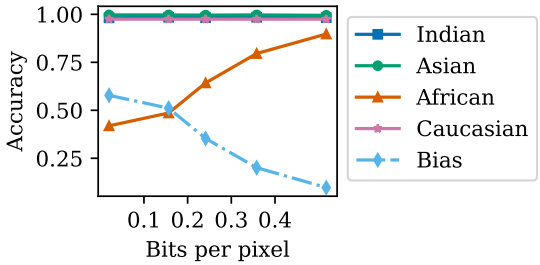

Figure 4: Bias for *skin type* across different races for *Joint* reconstructions trained on the CelebA dataset. As bitrate is lowered, the classification accuracy for the African group reduces significantly, leading to an increase of bias at the lower rates.

In this section, we evaluate bias in phenotype preservation with the metric defined in Equation 3. We illustrate this bias for one phenotype in Figure 4. In this figure, we observe, a significant decrease in the *skin type* classification accuracy for individuals in the African group at low bitrates for the *Joint* model. The accuracy for images in other racial groups have minimal changes. This reduction in accuracy for the African group leads to an amplification of bias as the bitrate is lowered.

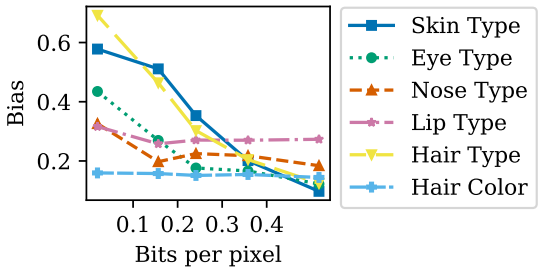

Figure 5: Bias across different phenotypes for *Joint* trained on the CelebA dataset. As the bitrate is lowered, bias increases for *Skin Type*, *Eye Type*, and *Hair Type*, while remaining relatively level for other phenotypes.

To better understand the amplification of bias for the *Joint* model, we highlight the bias across all phenotype tasks in Figure 5. Here, we observe that *skin type, eye type, and hair type* experience amplification of bias as we reduce bitrates while *nose type, lip type and hair color* do not. Specifically, the increase in bias is due to the reduction in accuracy for the the African group in hair type classification, and the non-Asian groups for eye type classification at lower rates.

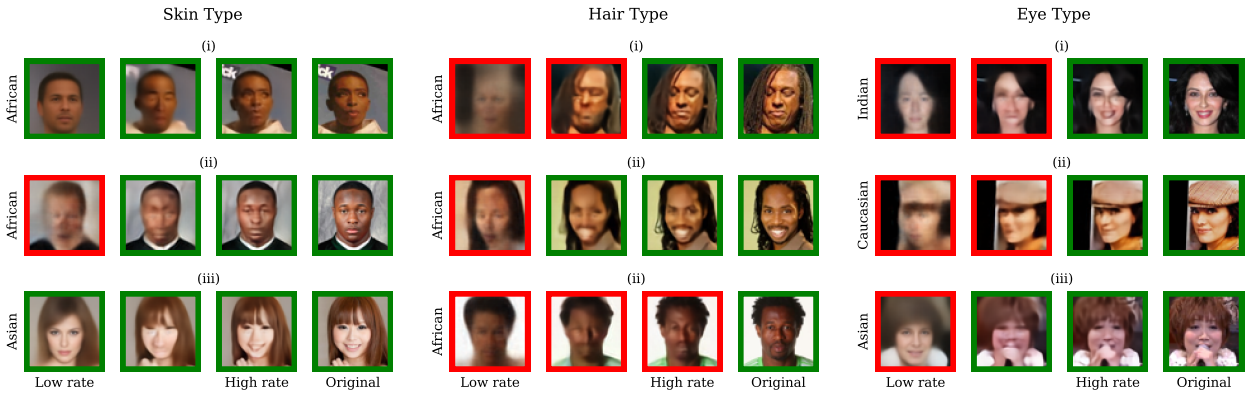

Figure 6: Image reconstructions from *Hyperprior* unless otherwise mentioned. **(a):** (i): African skin color lightens but is not captured by the classifier [*QRes*]. (ii): African skin color lightens and is captured by the classifier. (iii): Asian skin color changes but is not captured by the classifier [*GaussianMix-Attn*]. **(b):** (i): African curly hair becomes straight [*Joint*]. (ii-iii): African curly hair becomes straight. **(c)**: (i): Indian eye type becomes narrow. (ii): Caucasian eye type becomes narrow (iii): Asian eye type becomes normal [*GaussianMix-Attn*].

These plots and plots for other neural compression models are displayed in Appendix C.

The trends that we observe for the *Joint* model hold consistently across all neural compression architectures. The exception to this trend is the eye type classification task for the *GaussianMix-Attn* model. In this scenario, we observe an increase in bias due to the accuracy drop for the Asian group. We further investigate this in result in Section 4.3.

4.3. Qualitative Analysis of Image Reconstructions

We provide selected examples in Figure 6 and describe our main conclusions from a qualitative analysis of the image reconstructions across different bitrates. First, we observe that in most cases, individuals with darker skin have their skin lightened at lower bitrates. Through this effect we observe individuals for the African, Asian, and Indian groups appearing Caucasian at the lowest bitrates. This suggests that a "White Obama" problem occurs in neural compression at low bitrates. These conclusions are consistent with Jalal et al. (2021) where they find dark faces can be reconstructed to light faces after super-resolution. This phenomenon is typically captured by our phenotype classifier but occasionally missed. We demonstrate examples of this in Figure 6(a) (i) and 6a (iii). Additionally we observe the smoothing of hair type and eyes becoming more narrow as the bitrates are reduced. This is highlighted in Figure 6(b) and Figure 6(c). We conjecture that these effects are due to the natural process of neural compression which tends to blur images more at lower bitrates. This leads to the loss of features that are smaller and more correlated with textures in the original image. The one exception to this is in *GaussianMix-Attn*, where at the lowest rate, Asian eye types become normal (Figure 6(c) (iii)). This observation is consistent with the observation that the *GaussianMix-Attn* model produces more generic Caucasian faces at the lowest bitrates.

4.4. Balanced Dataset Comparison with FairFace

Next, we evaluate the effect of a balanced training dataset for neural compression models on the phenotype degradation bias. We use the FairFace dataset to train our neural compression models and examine how classification performance and bias changes as we reduce bitrates.

We highlight the results of this experiment for the *GaussianMix-Attn* architecture in Figure 7. The dataset comparison plots slightly vary across other neural compression architectures (Appendix D). We observe, in general, that for skin type, eye type, and hair type, the FairFace dataset has minimal effect in reducing bias and in many cases amplifies bias. This suggests that training on a race-balanced dataset is not sufficient to resolve bias for low-rate compression. This finding is consistent with that of Cherepanova et al. (2023) that class-balanced learning does not necessarily lead to fair classification. Additionally, the amplification of bias by using FairFace could be attributed to the facial orientation differences between FairFace and CelebA (Laszkiewicz et al., 2024). Facial images in CelebA are mostly forward facing, while facial orientations in FairFace are more diverse. This difference in data quality could make it easier for models to learn an average face from CelebA than from FairFace. For future works, it will be beneficial to train neural compression models on two datasets that have similar face orientations.

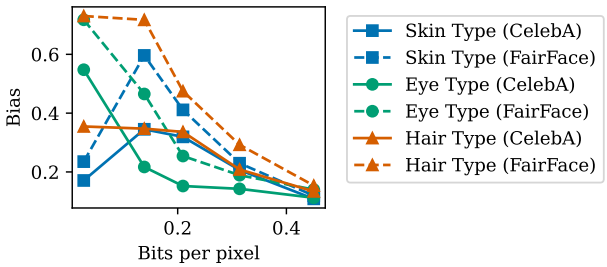

Figure 7: Bias for Skin Type, Eye Type, and Hair Type for the *GaussianMix-Attn* model. A racially balanced training set (FairFace) does not generally improve bias compared to an imbalanced training set (CelebA).

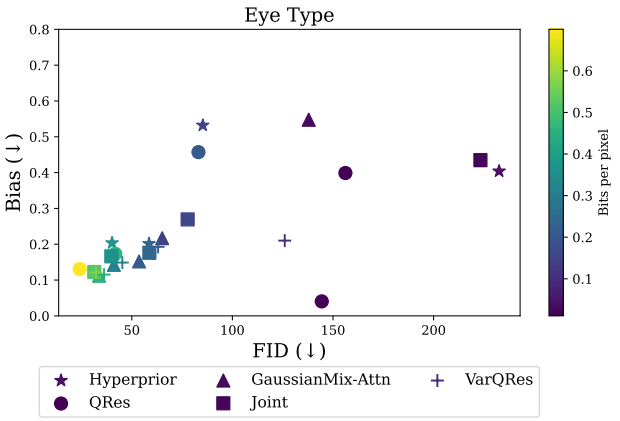

Figure 8: At high bitrates (> 0.1 bpp), there is a strong linear correlation between bias and realism. At low bitrates (< 0.1 bpp), however, this correlation diminishes and the relationship between bias and realism is more variable.

4.5. Bias-Realism Relationship

We systematically assess the relationship between bias and realism across various neural compression models. Realism is quantified using FID (Heusel et al., 2017) and bias values are derived from Equation 3. To ensure that we are measuring realism with respect to general facial datasets, we utilize the Demogpairs dataset (Hupont & Fernández, 2019) as the reference dataset for computing FID. This approach enables us to capture the fidelity of the reconstructions without spurious correlations.

The relationship between bias and realism is highlighted in Figure 8. Overall, for all phenotypes that show significant bias (i.e., skin type, hair type, eye type), we observe a linear correlation between bias and realism at higher rates (larger than 0.1 bpp). This indicates that models that produce highly biased reconstructions also produce unrealistic reconstructions. However, at low bit rates (less than 0.1 bpp), this relationship becomes more sporadic. We believe that the exploration of the relationship between bias and realism at low bit rates remains an interesting future direction. Further work in neural compression can explore the development of models that produce values with high realism and low bias at low bitrates.

5. Conclusion and Discussion

We reveal bias in phenotype loss under low-rate neural compression, notably for non-Asian individuals' eye types and African individuals' skin and hair type. We find that racially balancing training data fails to mitigate this bias. Furthermore, at low bitrates, the relationship between realism and bias becomes more variable. This pioneering analysis of bias in low-rate neural image compression prompts further exploration. Future research directions include:

Phenotype classifier variability While we report the results for a single classifier, we have observed that per-race accuracy and bias can vary significantly with different random initialization of the classifier while the overall accuracy remains mostly constant. We relate this to predictive multiplicity (Hsu & Calmon, 2022; Marx et al., 2020) and understanding the variability of the classifier should be further investigated.

Isolating bias For evaluation, we utilize a single phenotype classifier across different bitrates. This allows us to isolate the bias of the classifier by examining the performance differences across difference rates. Future work can further investigate isolating the bias of the phenotype classifier by leveraging a fair classifier. Dooley et al. (2023) demonstrate that bias can be inherent to the classifier architecture and that fair architectures can be found through neural architecture search. Exploring a fair architecture for neural compression is an interesting future direction. Additionally, emerging information theoretic techniques (Goldfeld & Greenewald, 2021; Goldfeld et al., 2022; Wongso et al., 2022; 2023; Tax et al., 2017; Wibral et al., 2017; Dutta et al., 2020; Dutta & Hamman, 2023) can be explored to further decouple bias in the encoder and decoder of neural compression architectures.

Bias in realism-oriented compression Some neural compression models incorporate an adversarial loss term to increase the realism of decoded images (Agustsson et al., 2023). It is under-explored if there is a trade-off between the realism and the fairness.

Bias in semantic compression Powerful foundation models (Radford et al., 2021) give rise to "zero-shot" semantic image compressors, where the encoder is an image-to-text model, and decoder a text-to-image model (Lei et al., 2023). While there is research that tackles the bias in foundation models (Abid et al., 2021; Bommasani et al., 2021), little work has been done to investigate these semantic image compressors.

Impact Statement

This paper presents work whose goal is to examine bias present in popular, deep learning-based, neural compression algorithms. Bias in neural compression can adversely affect downstream tasks such as biometric analysis. Our goal is to promote downstream fairness across all tasks which may stem from neurally compressed data. The removal of bias has positive societal impacts, as we strive towards building fair AI systems.

References

Abid, A., Farooqi, M., and Zou, J. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Agustsson, E., Minnen, D., Toderici, G., and Mentzer, F. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22324–22333, 2023.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Bégaint, J., Racapé, F., Feltman, S., and Pushparaja, A. CompressAI: a PyTorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

Bellard, F. BPG image format, 2014. URL https://bellard.org/bpg/. Accessed: 2024-05-24.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art, 2017.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bross, B., Wang, Y.-K., Ye, Y., Liu, S., Chen, J., Sullivan, G. J., and Ohm, J.-R. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31 (10):3736–3764, 2021.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Chen, B., Yin, S., Chen, P., Wang, S., and Ye, Y. Generative visual compression: A review. *arXiv preprint arXiv:2402.02140*, 2024.

Cheng, Z., Sun, H., Takeuchi, M., and Katto, J. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7939–7948, 2020.

Cherepanova, V., Reich, S., Dooley, S., Souri, H., Dickerson, J., Goldblum, M., and Goldstein, T. A deep dive into dataset imbalance and bias in face identification. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pp. 229–247, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604691. URL https://doi.org/10.1145/3600211.3604691.

Dooley, S., Sukthanker, R., Dickerson, J., White, C., Hutter, F., and Goldblum, M. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74366–74393. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/eb3c42ddfa16d8421fdba13528107cc1-Paper-Conference.pdf.

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., and Busch, C. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

Duan, Z., Lu, M., Ma, J., Huang, Y., Ma, Z., and Zhu, F. Qarv: Quantization-aware resnet vae for lossy image compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023a.

Duan, Z., Lu, M., Ma, Z., and Zhu, F. Lossy image compression with quantized hierarchical vaes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 198–207, 2023b.

Dutta, S. and Hamman, F. A review of partial information decomposition in algorithmic fairness and explainability. *Entropy*, 25(5):795, 2023.

Dutta, S., Venkatesh, P., Mardziel, P., Datta, A., and Grover, P. An information-theoretic quantification of discrimination with exempt features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3825–3833, 2020.

Ez-Zazi, I., Arioua, M., and El Oualkadi, A. Analysis of lossy compression and channel coding tradeoff for energy efficient transmission in low power communication systems. In *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pp. 291–295. IEEE, 2018.

Gao, F., Deng, X., Jing, J., Zou, X., and Xu, M. Extremely low bit-rate image compression via invertible image generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Goldfeld, Z. and Greenewald, K. Sliced mutual information: A scalable measure of statistical dependence. *Advances in Neural Information Processing Systems*, 34:17567–17578, 2021.

Goldfeld, Z., Greenewald, K., Nuradha, T., and Reeves, G. k-sliced mutual information: A quantitative study of scalability with dimension. *Advances in Neural Information Processing Systems*, 35:15982–15995, 2022.

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

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hsu, H. and Calmon, F. Rashomon capacity: A metric for predictive multiplicity in classification. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 28988–29000. Curran Associates, Inc., 2022.

Hu, S. and Chen, W. Joint lossy compression and power allocation in low latency wireless communications for iiot: A cross-layer approach. *IEEE Transactions on Communications*, 69(8):5106–5120, 2021.

Hupont, I. and Fernández, C. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pp. 1–7. IEEE, 2019.

Jalal, A., Karmalkar, S., Hoffmann, J., Dimakis, A., and Price, E. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, pp. 4721–4732. PMLR, 2021.

Kärkkäinen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., and Jain, A. K. Face recognition performance: Role of demographic information. *IEEE Transactions on information forensics and security*, 7(6):1789–1801, 2012.

Laszkiewicz, M., Daunhawer, I., Vogt, J. E., Fischer, A., and Lederer, J. Benchmarking the fairness of image upsampling methods. *arXiv preprint arXiv:2401.13555*, 2024.

Lei, E., Uslu, Y. B., Hassani, H., and Bidokhti, S. S. Text+ sketch: Image compression at ultra low rates. *arXiv preprint arXiv:2307.01944*, 2023.

Li, M., Shen, L., Ye, P., Feng, G., and Wang, Z. Rfd-ecnet: Extreme underwater image compression with reference to feature dictionary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12980–12989, 2023.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15 (2018):11, 2018.

Marx, C., Calmon, F., and Ustun, B. Predictive multiplicity in classification. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6765–6774. PMLR, 13–18 Jul 2020.

Minnen, D., Ballé, J., and Toderici, G. D. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018.

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

Serna, I., Morales, A., Fierrez, J., Cebrian, M., Obradovich, N., and Rahwan, I. Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics. *arXiv preprint arXiv:1912.01842*, 2019.

Tanjim, M. M., Singh, K. K., Kafle, K., Sinha, R., and Cottrell, G. Debiasing image-to-image translation models. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. URL https://bmvc2022.mpi-inf.mpg.de/0182.pdf.

Tax, T. M., Mediano, P. A., and Shanahan, M. The partial information decomposition of generative neural network models. *Entropy*, 19(9):474, 2017.

Toderici, G., O'Malley, S. M., Hwang, S. J., Vincent, D., Minnen, D., Baluja, S., Covell, M., and Sukthankar, R. Variable rate image compression with recurrent neural networks. *arXiv preprint arXiv:1511.06085*, 2015.

Toderici, G., Vincent, D., Johnston, N., Jin Hwang, S., Minnen, D., Shor, J., and Covell, M. Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5306–5314, 2017.

Townsend, J., Bird, T., and Barber, D. Practical lossless compression with latent variables using bits back coding. *arXiv preprint arXiv:1901.04866*, 2019.

Vangara, K., King, M. C., Albiero, V., Bowyer, K., et al. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Wallace, G. K. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 692–702, 2019.

Wibral, M., Finn, C., Wollstadt, P., Lizier, J. T., and Priesemann, V. Quantifying information modification in developing neural networks via partial information decomposition. *Entropy*, 19(9):494, 2017.

Wongso, S., Ghosh, R., and Motani, M. Understanding deep neural networks using sliced mutual information. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 133–138. IEEE, 2022.

Wongso, S., Ghosh, R., and Motani, M. Using sliced mutual information to study memorization and generalization in deep neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 11608–11629. PMLR, 2023.

Xu, T., White, J., Kalkan, S., and Gunes, H. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 506–523. Springer, 2020.

Yang, Y., Mandt, S., Theis, L., et al. An introduction to neural data compression. *Foundations and Trends® in Computer Graphics and Vision*, 15(2):113–200, 2023.

Yucer, S., Poyser, M., Al Moubayed, N., and Breckon, T. P. Does lossy image compression affect racial bias within face recognition? In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10. IEEE, 2022a.

Yucer, S., Tektas, F., Al Moubayed, N., and Breckon, T. P. Measuring hidden bias within face recognition via racial phenotypes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 995–1004, 2022b.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

A. Training details

A.1. Phenotype Classifier

We fine-tuned pretrained ResNet18 models (He et al., 2016) for facial phenotype classification. The classifiers retain the ResNet18 backbone and include a classification head for classifying the specific attribute. We trained the separate phenotype classifier models for up to 20 epochs, employing early stopping with patience of 5 epochs. We use cross entropy loss and optimize the models with the stochastic gradient descent optimizer, a fixed learning rate of 0.01, and a fixed batch size of 32. To evaluate each compression model at different compression rates, we train the models on decompressed images from each of the evaluated neural compression models with different compression rates separately, using the provided dataset annotations.

A.2. Neural Compression Models

For the CompressAI neural compression models, we train for 1000 epochs with an early stopping patience of 50 epochs. We use a batch size of 64 and an initial learning rate of 0.0001. For the rest of the parameters, we leave them as they are implemented in the CompressAI repository. For the *QRes* (Duan et al., 2023b) and *VarQres* (Duan et al., 2023a) implementations, we follow the training procedure from the papers.

B. Phenotype Degradation

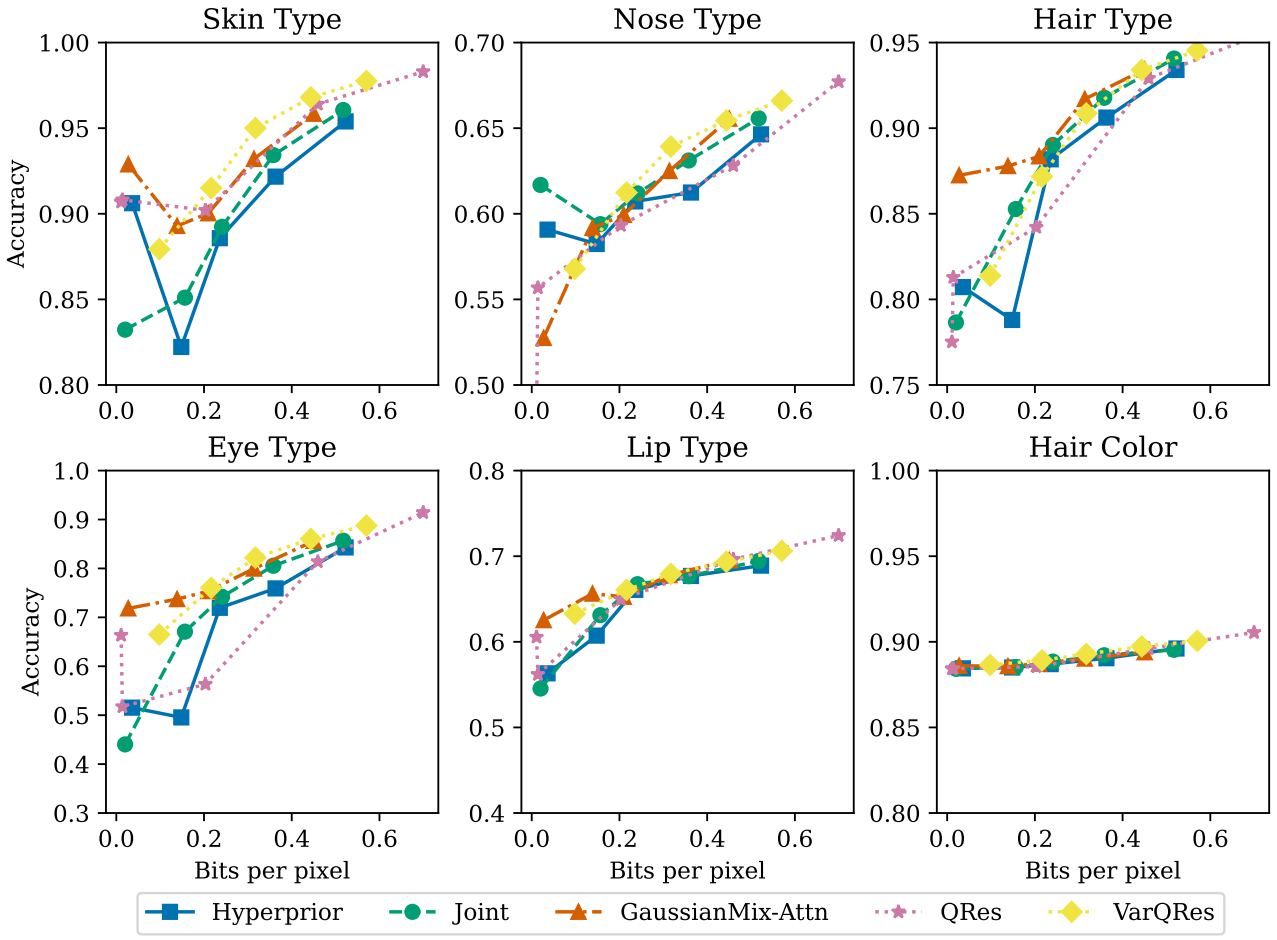

Figure B.1: Phenotype degradation across all compression models for the CelebA dataset.

C. Racial Bias in Degradation

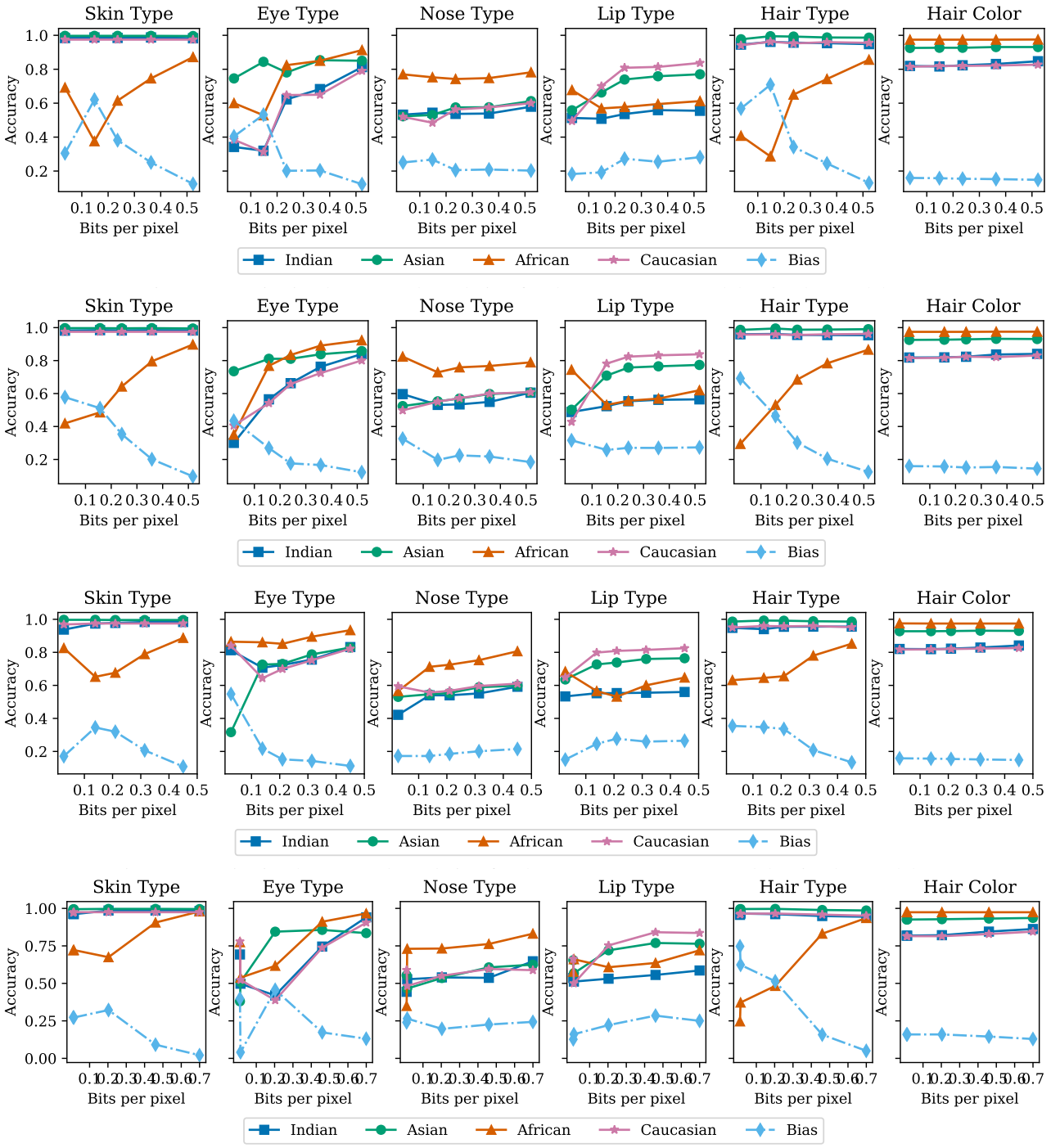

Figure C.4: Bias in phenotype degradation for the *QRes* Model trained on CelebA

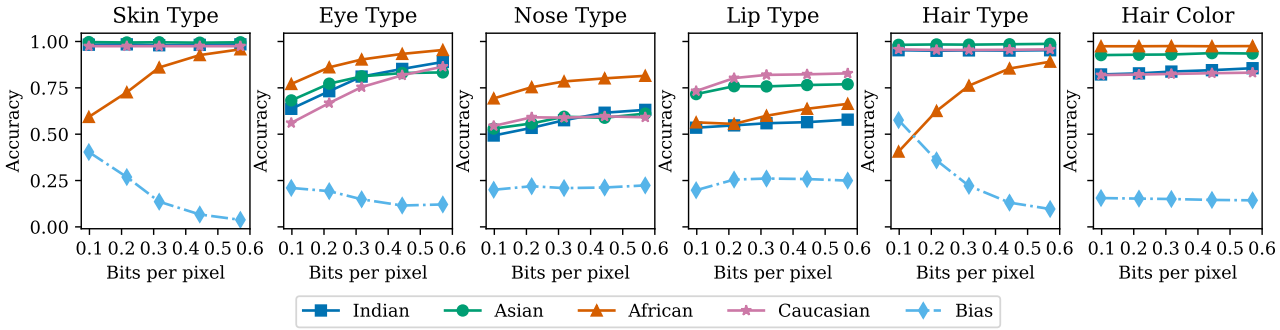

Figure C.5: Bias in phenotype degradation for the *VarQRes* Model trained on CelebA

D. Training with a Balanced Dataset

In Figure D we present the impact of using a balanced training set FairFace on racial bias in phenotype degradation.

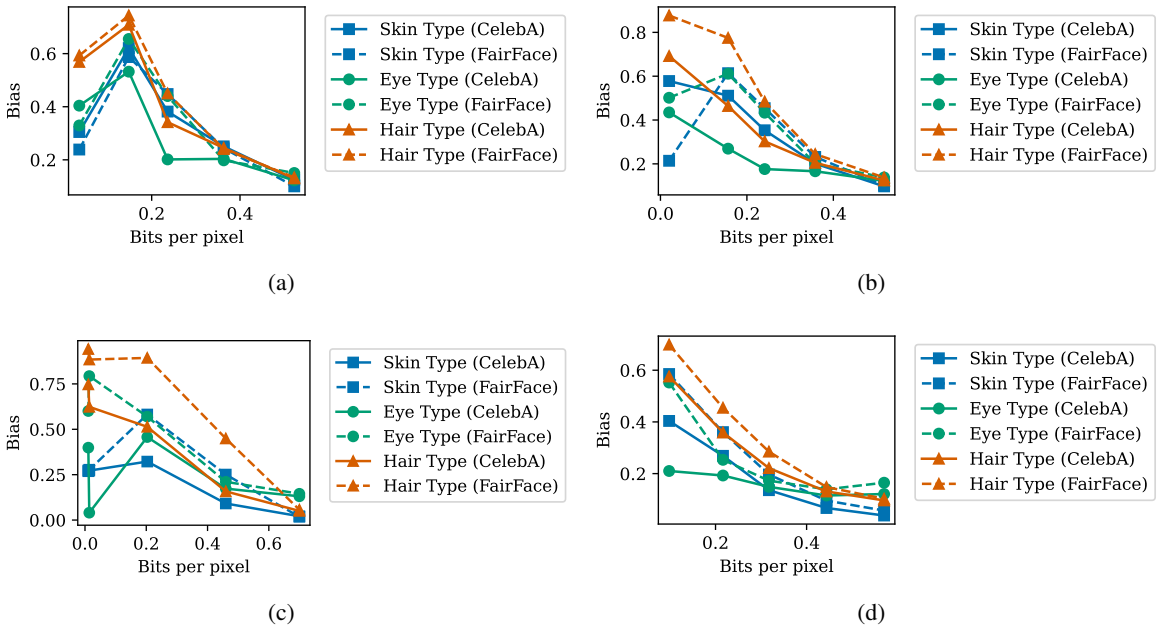

Figure D.1: (a) Hyperprior (b) Joint (c) QRes (d) VarQRes

E. Bias-Realism Trade-off

In Figure E we present FID vs bias figures for all the phenotypes.

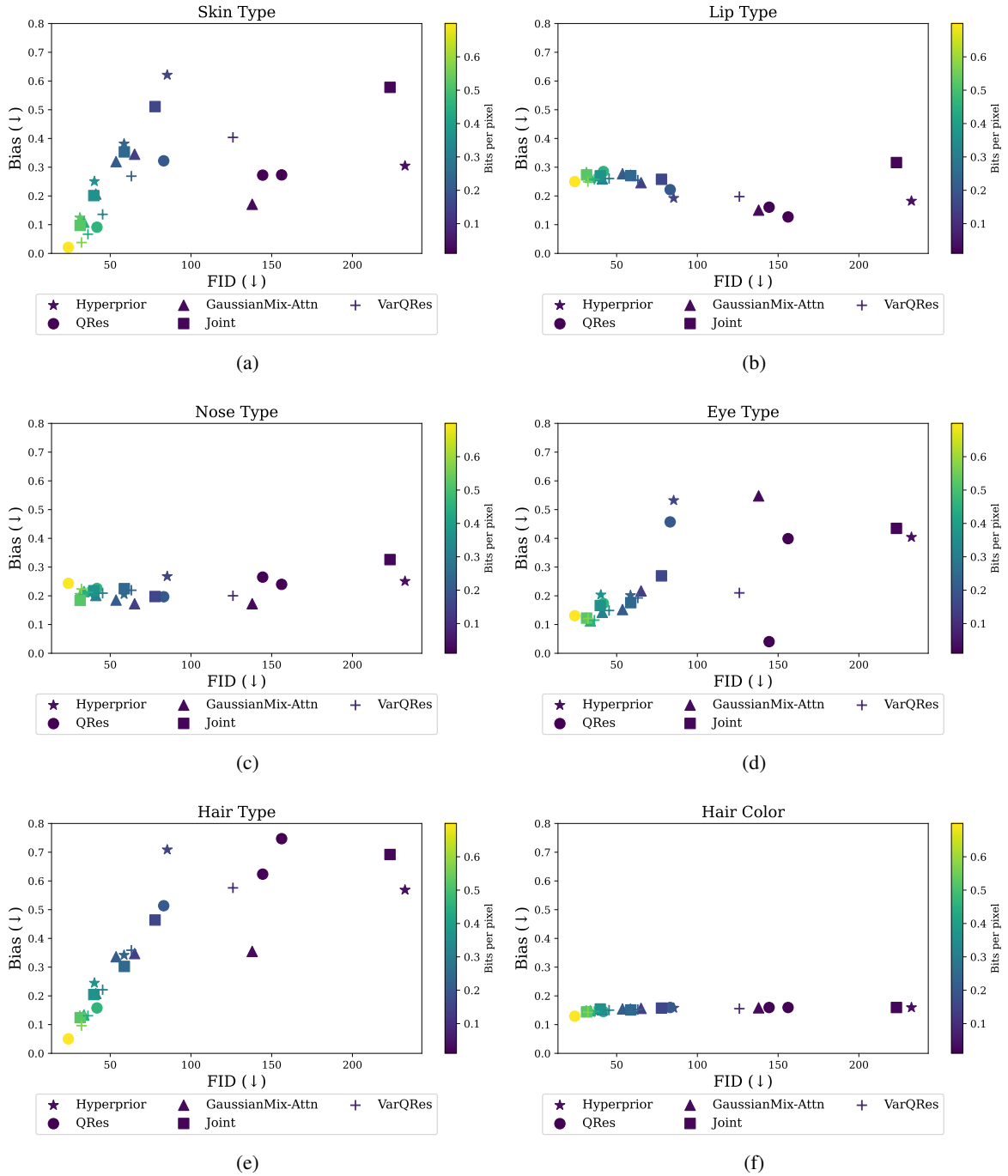

Figure E.1: (a) Skin Type (b) Lip Type (c) Nose Type (d) Eye Type (e) Hair Type (f) Hair Color