Debiasing the Denoising U-Net: Mitigating Bias in Stable Diffusion With LoRA

Reece Iriye
Southern Methodist University
Department of Mathematics
Dallas, TX, US
ririye@smu.edu

Leland Winston
Southern Methodist University
Department of Computer Science
Dallas, TX, US
lwinston@smu.edu

Trevor Dohm
Southern Methodist University
Department of Computer Science
Dallas, TX, US
tdohm@smu.edu

Kassi Bertrand
Southern Methodist University
Department of Computer Science
Dallas, TX, US
knzalasse@smu.edu

Eric Larson
Southern Methodist University
Department of Computer Science
Dallas, TX, US
elarson@smu.edu

Abstract

Generative models like Stable Diffusion have revolutionized image generation by enabling the creation of highly realistic images from textual descriptions. However, these models can inadvertently absorb and perpetuate societal biases present in the training data, leading to the generation of images that reinforce stereotypes. We explore the mitigation of intersectional bias on the basis of race and gender in image generation by training low-rank decomposition matrices and applying Low-Rank Adaptation (LoRA) to effectively "debias" the weights in the cross-attention heads of each U-Net on each iteration through the denoising process. To address these biases with LoRA, we generate training data using a diffusion model with equally distributed combinations of race and gender across a large breadth of designations/labels/nouns that can be assigned to people. Then, we perform our novel approach, Caption Excising, to effectively remove the explicit race and gender information from the prompts and assign the remaining prompts as captions to the images. We hypothesize that the LoRA will prevent the absorption of biases related to race and gender, and instead will inject a tendency to associate designations with a large breadth of intersectional identities. We compare the frequency and distribution of race and gender in images generated from prompts that do not explicitly mention race or gender, using both the original Stable Diffusion model and the model with the LoRA applied. We find that there is a clearly notable difference in race distributions and an especially notable difference in gender distributions varying by designation for each prompt.[1]

1 Introduction

Diffusion models [15] have revolutionized image synthesis, enabling the creation of highly realistic images from textual descriptions. Among these models, Stable Diffusion has emerged as the preferred architecture for image generation tasks, effectively ending the long-standing dominance of Generative Adversarial Networks (GANs) [4]. Stable Diffusion models learn from vast amounts of data, capturing the patterns and relationships in the training dataset [16].

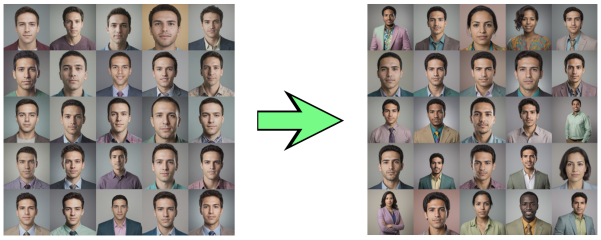

Figure 1. Samples of 25 images generated with the prompt "An individual successful person, generated in full color, facing towards the camera." by RealVisXL v4.0 before & after fusing our trained LoRA with scale factor 0.6.

However, despite their impressive capabilities, Stable Diffusion models can inadvertently perpetuate the biases present in the training data [2, 20]. These biases can manifest in various forms, such as the underrepresentation or misrepresentation of certain demographics, the reinforcement of gender roles, or the association of specific attributes with particular identities [10]. As a result, images generated using Stable Diffusion may reinforce stereotypes and lack diversity, leading to significant societal implications by shaping public perception and reinforcing existing inequalities.

Motivated by these concerns, our research aims to investigate the mitigation of intersectional bias related to race and gender in Stable Diffusion models. We seek to answer the following research questions:

1. To what extent can Low-Rank Adaptation (LoRA) be used to mitigate racial and gender biases in Stable Diffusion models?
2. How effective is our novel approach of Caption Excision in creating a balanced dataset for training the LoRA matrices?

[1]Code: https://github.com/reece-iriye/Mitigating-Bias-in-Stable-Diffusion-Models-Using-LoRA

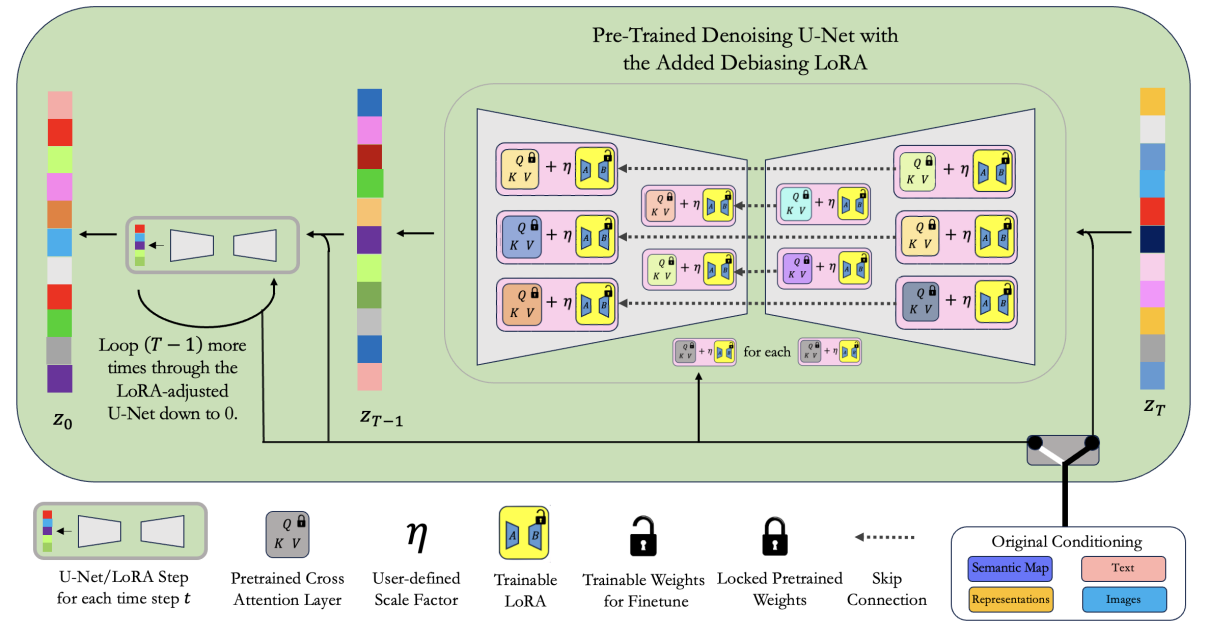

Figure 2. LoRA applied to the cross-attention layers of each denoising U-Net in the diffusion model.

3. What impact does our bias mitigation technique have on the frequency and distribution of race and gender in images generated from prompts that do not explicitly mention these attributes?

By addressing these questions, we aim to contribute to the development of more equitable and inclusive generative models that better reflect the diversity of our society. Our research has the potential to provide valuable insights and techniques for mitigating bias in Stable Diffusion and other generative models, thereby promoting fair representation and reducing the reinforcement of stereotypes.

Our approach involves training low-rank decomposition matrices and applying Low-Rank Adaptation (LoRA) [7] to adjust the frozen cross-attention weights in the U-Net [5, 12] on each iteration of the diffusion process [15]. By training the LoRA matrices, we aim to steer the diffusion model away from its biased tendencies. We leverage the practicality and effectiveness of LoRA, as it has the ability to adjust the style and transform the weights in the model, enabling specialization and generalization in a desired direction [7]. The LoRA can be seen in the latent space defined by Figure 2.

To train the LoRA matrices, we generate a balanced dataset using a diffusion model with equally distributed combinations of race and gender across a wide range of labels and nouns, referred to as "designations" throughout the paper. We then employ a novel approach called Caption Excision, which removes the explicit race and gender information from the prompts and assigns the resulting prompts as captions to the images they originally generated. By training the LoRA on this balanced and caption-excised dataset, we hypothesize that the model will learn to associate designations with a diverse range of intersectional identities, rather than solely representing the potentially biased data on which the Stable Diffusion model was originally trained.

To evaluate the effectiveness of our approach, we compare the frequency and distribution of race and gender in images generated from prompts that do not explicitly mention race or gender. We generate these images using both the original Stable Diffusion model and the model with LoRA, allowing us to assess the impact of our bias mitigation technique.

2 Related Work

2.1 Bias in Generative Models

Bias in generative models has been a topic of growing concern in the research community. Several studies have investigated the presence and impact of biases in various domains of machine learning, including natural language processing [9], computer vision [20], and multimodal learning [2].

Stable Diffusion [15] has been shown to exhibit biases in its generated images. These biases can be attributed to the training data, which often contains societal biases and stereotypes [16]. Additionally, the integration of text and image embeddings through CLIP [14] can further reinforce these biases, which has been addressed previously [19].

2.2 Reducing Bias in Stable Diffusion

Efforts to mitigate biases in generative models have explored various approaches. Data curation methods, for example, aim to create more balanced and diverse datasets by carefully selecting or augmenting the training data [1]. These curated datasets are extremely useful; however, training a diffusion model from scatch requires extensive computing power that

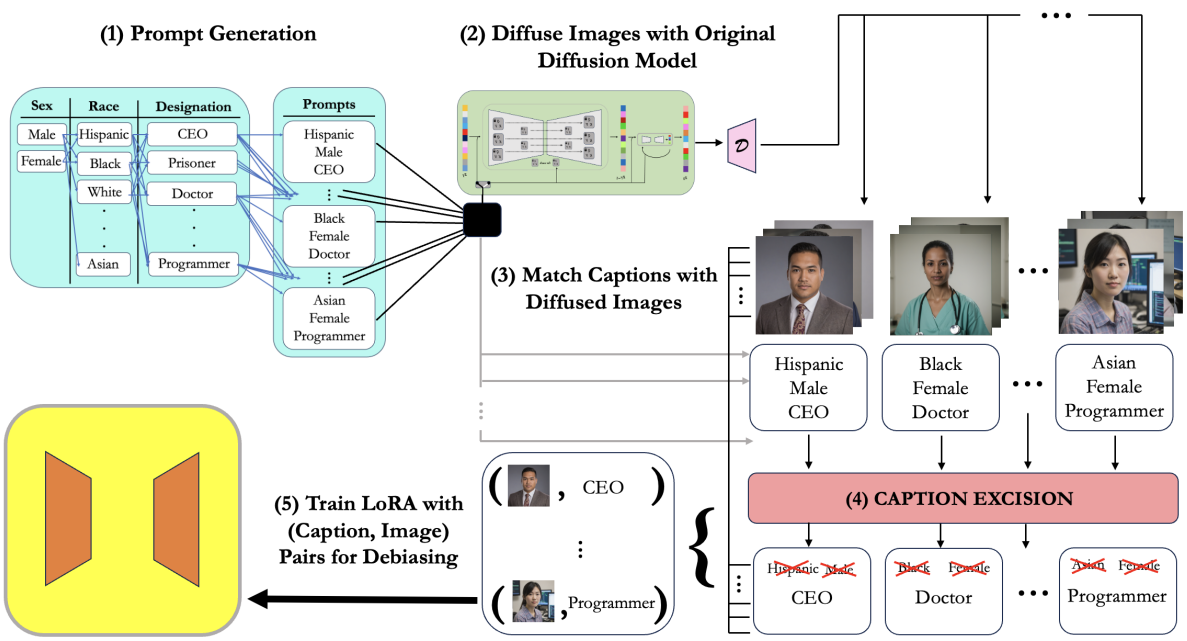

Figure 3. Overview of the data generation process for creating a diverse dataset encompassing various combinations of race, gender, and designations to train the LoRA model.

most people cannot access [15]. We employ learned lessons about dataset curation into our LoRA-training experiment.

Model-based approaches, such as editing implicit assumptions in cross-attention layers [3, 10] or employing disentangled representations [11], have shown some promise in reducing biases. These methods, however, often require modifications to the model architecture or training objectives and are too expensive to be employed universally.

2.3 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation (LoRA) [7] has emerged as a parameter-efficient fine-tuning technique for adapting pre-trained models to new tasks or domains. LoRA has been successfully applied in various natural language processing tasks, such as dialogue systems [6], text classification [8], and machine translation [21]. However, the application of LoRA for bias mitigation in generative models remains unexplored.

We leverage the advantages of LoRA to mitigate biases on the basis of race and gender in Stable Diffusion. We hypothesize that by fine-tuning a Stable Diffusion model on a balanced dataset using LoRA, we can override previously obtained associations between race/gender to some extent. As a result, we would achieve a more equitable representation in the generated images while minimizing the computational overhead associated with traditional fine-tuning.

3 Methodology

3.1 Selecting a Stable Diffusion Model

We chose RealVisXL v4.0[2] as our Stable Diffusion model for this study due to its ability to generate highly realistic images and its consistency in producing real-looking images of people. RealVisXL v4.0 is a fine-tuned version of SDXL [13], which already has an impressive capacity to generate realistic images. By using RealVisXL v4.0, we minimize the influence of stylistic variations on our analysis, allowing us to focus on the impact of our bias mitigation approach. The model's consistency in generating realistic images of people across different prompts ensures that any observed changes in the distribution of race and gender can mostly be attributed to the effectiveness of our methodology, rather than inherent variations in the model's output style.

3.2 Training Data Creation

The purpose of this LoRA is to inject debiasing behavior into the cross-attention heads of the U-Net within the Stable Diffusion model. To effectively train the LoRA matrices, it is crucial to ensure that the dataset has a balanced representation of race and gender, taking into account intersectionality [1]. In order to meet this requirement, we curated a custom dataset using the original model. The images in this dataset were generated using a specific prompt structure:

[2]RealVisXL v4.0 HuggingFace Model Repository: https://huggingface.co/SG161222/RealVisXL_V4.0

An individual *(gender) (race) (designation)* generated in full color, facing towards the camera.

This prompt structure simplifies the process of evaluating the model's performance by providing a consistent framework for generating images across different combinations of gender, race, and designation. By using this standardized prompt, we can more easily assess the impact of our debiasing methodology on the generated images.

The dataset is structured to cover a wide range of identities and attributes. We include 34 distinct designations, representing various roles and identities that a person can hold. To capture global diversity, we incorporate 18 different races/ethnicities in our dataset. Additionally, we consider two genders (male and female) in our data generation process. By combining these attributes, we generate a total of 1,224 unique prompts. For each prompt, we use RealVisXL v4.0 to generate 8 images, resulting in a comprehensive training dataset of 9,792 images with associated captions.[3] This process is illustrated in Steps 1, 2, & 3 of Figure 3. All races and designations used for the training prompts can be found in Appendix A.1 and Appendix A.2, respectively.

To create captions for each image, we employ Caption Excision, as outlined in Step 4 of Figure 3. To evaluate how the training data might influence the LoRA, we analyze some summary characteristics of the dataset. Using the DeepFace Python library [17], we detect faces in the generated images. For each detected face, we extract the forehead region and calculate the average hue of the pixels in that region to represent the overall hue of the face. This hue value serves as a proxy for the skin tone of the individual in the image.

By examining the distribution of hues across the dataset, we observe that the generated faces tend to have darker hues, as shown in Figure 4. This suggests that the model may be generating a higher proportion of individuals with darker skin tones. However, it is important to note that the hue distribution could be influenced by factors beyond just skin tone, such as lighting conditions or other image characteristics, which could lead to some issues.

Despite this potential bias in the hue distribution, our training dataset is designed to have an equal representation of genders, with 50% male and 50% female images. This balanced gender distribution is ensured by the prompt structure used to generate the images, which explicitly includes both male and female designations.

3.3 LoRA Training Methodology

Our approach to mitigating biases in Stable Diffusion models involves training a LoRA model using the generated training data with excised captions. LoRA is a parameter-efficient fine-tuning technique that adapts pre-trained models to new tasks

[3]HuggingFace Dataset Repository for LoRA Training Images: https://huggingface.co/datasets/ririye/Generated-LoRA-Input-Images-for-Mitigating-Bias

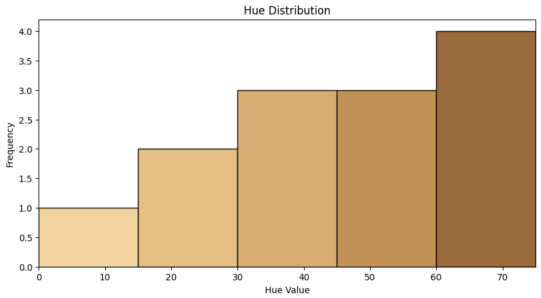

Figure 4. Foreheads hues of generated training data

by adding a small number of trainable parameters [7]. By targeting the cross-attention layers in the U-Net architecture of the Stable Diffusion model, we aim to directly influence the model's ability to generate images that are less biased towards specific racial and gender representations.

One of the key advantages of using LoRA for bias mitigation is its computational efficiency compared to training a diffusion model from scratch. Training a complete diffusion model requires a significant amount of computational resources and time, making it infeasible for most researchers and practitioners. In contrast, LoRA allows for effective fine-tuning of pre-trained models with limited computational power. By adding only a small number of trainable parameters to the existing model, LoRA significantly reduces the computational burden while still achieving the desired bias mitigation effects. This makes our approach more accessible and practical for a wider range of users, even those without access to high-end GPU resources.

To ensure that the LoRA training focuses on the cross-attention layers, we train only the U-Net while keeping the other components frozen. This is particularly important when training LoRAs for SDXL, which contains dual text encoders. Training both text encoders simultaneously can lead to problematic results, so we specifically target the U-Net and cache the text encoder outputs during training.

Another advantage of our LoRA training methodology is its flexibility in terms of the base model. Any Stable Diffusion model can be used as the starting point, regardless of the base training method or its specific behavior, since we target the cross-attention layers. This allows for a wide range of applications and adaptability to different model architectures. However, it is important to note that if the base model has a Variational Autoencoder (VAE) baked into it, some preprocessing may be required to extract the VAE before performing the LoRA training to avoid potential issues.

The hyperparameters used for training the LoRA play a crucial role in determining the effectiveness of the bias mitigation. We carefully select and tune these hyperparameters based on empirical observations and prior work on LoRA training for Stable Diffusion models. The specific values these hyperparameters take will be provided in the results section,

along with a discussion on their impact on the training process and the generated outputs, as well as other information such as the optimizer, network dimension and alpha, which are are important factors in controlling the size and capacity of the LoRA layers. These parameters are chosen to strike a balance between expressiveness and efficiency, allowing the model to learn more balanced representations while maintaining computational feasibility.

4 Results

In this section, we present the results of our LoRA training methodology for bias mitigation in Stable Diffusion models. We evaluate the effectiveness of our approach by comparing the generated images from the base model and the fine-tuned model using various metrics and qualitative analyses.

4.1 Experimental Setup

For our experiments, we used the RealVisXL v4.0 model as the base Stable Diffusion model. We trained the LoRA using the generated training data with excised captions, as described in the methodology section. The training was conducted using the following hyperparameters: a learning rate of 5e-06, a batch size of 32, and a total of 30 epochs. We employed the Adafactor optimizer and used mixed precision (fp16) to optimize memory usage and training speed. Gradient accumulation steps were set to 1, and gradient clipping was applied with a value of 1.0. The network dimension and alpha were set to 64 and 32, respectively, and L2 loss was used as the training objective.

The training was performed on a system equipped with 8 NVIDIA A100 GPUs, leveraging data parallelism to distribute the workload across multiple devices. The total training time amounted to approximately 24 hours.[4]

After training the LoRA, we generated 1,000 images for each of the 12 test labels (Appendix A.3) using the following prompt format:

An individual *(designation)* generated in full color, facing towards the camera.

These test labels include a mix of designations that were present in the LoRA training captions and some that were not. This allows us to evaluate the effectiveness of the LoRA in mitigating biases for both seen and unseen designations. By generating a large number of images for each test label, we can obtain a robust assessment of the LoRA's impact on the generated images' diversity and fairness.[5,6]

[4]HuggingFace Debiased LoRA Model Repository: https://huggingface.co/ririye/LoRA-Debiased-RealVisXL_V4.0

[5]HuggingFace Dataset Repository for Pre-LoRA Images: https://huggingface.co/datasets/ririye/Benchmark-Images-for-Stable-Diffusion-Bias

[6]HuggingFace Dataset Repository for Post-LoRA Images: https://huggingface.co/datasets/ririye/Mitigated-Bias-Comparison-Images

4.2 Quantitative Analysis

To quantitatively assess the bias mitigation effects of our LoRA training, we generated bar plots that charted the frequency of gender and race appearance using the base model and the fine-tuned model for each of the 12 test labels (Appendix A.3). Along with these bar plots we also made histograms that display the frequency of average hue values in each individuals forehead. These prompts included various combinations of designations, such as "doctor," "activist," "CFO," etc., without specifying any racial or gender information. While some of the designations used were captions in the LoRA training process, most of them were not.

We then utilized the DeepFace library [17, 18] to analyze the generated images and extract demographic information, including perceived race, gender, and average hue. The results were aggregated and compared between the base model and the fine-tuned model.

Figures 5 - 9 (see Appendix B.1) illustrate the distribution of perceived race in the generated images for the base model. While there was a tendency for the model to generate images of individuals perceived as White, this bias was not consistent across all prompts. For example, the model disproportionately generated individuals classified as Middle Eastern for the "terrorist" designation and individuals classified as Hispanic - Latino for the "activist" designation.

Figures 10 - 14 (see Appendix B.2)) present the distribution of perceived gender in the generated images from the benchmark model. The model exhibited a notable bias towards generating images of individuals perceived as male across most designations.

Figures 15 - 19 (see Appendix B.3)) display the average hue distribution across the designations for the benchmark model. The model disproportionately generated lighter images for most prompts, except for the "terrorist" prompt.

For the post-LoRA fine-tuned model, the bar plots and histograms show that the LoRA affected the distribution of appearances, but the resulting distribution is not even and is unpredictable. These distributions do show a significant change in outcomes but could be due to various factors, including the limitations of the Deepface library [17] [18] or our method for extracting hue values from foreheads.

Please note that bar plots used in the analyses of race and gender classificatio are not to scale on the y-axis as they display the counts of race and gender occurrence.

4.3 Qualitative Analysis

In addition to the quantitative analysis, we conducted a qualitative assessment of the generated images to gain insights into the visual quality and diversity of the outputs. Figures 20 - 24 (see Appendix B.4) showcases a side-by-side comparison of images generated by the original model and the fine-tuned model for various prompts.

In our qualitative analysis we found that the fine-tuned model generated images with more diverse racial and gender representations while maintaining the overall quality and coherence of the generated images. The images produced by the fine-tuned model exhibited a wider range of skin tones, facial features, and gender.

These analyses show that our method of fine-tuning the diffusion model with a LoRA can significantly influence model output. However, it's important to note that racial classification is an ambiguous metric, and hue value might be a better measure for assessing skin tone diversity in generated images, as it is a more objective metric.

4.4 Limitations

While our LoRA training methodology demonstrates promising results in mitigating biases, there are certain limitations to be addressed in future work. One potential limitation is the reliance on the DeepFace library for demographic analysis, which may introduce its own biases and inaccuracies. Future research could explore alternative methods for assessing demographic attributes in generated images.

Another area for future investigation is the generalization of our approach to other generative models and domains. While we focused on Stable Diffusion models for image generation, the principles of LoRA training for bias mitigation could be extended to other modalities, such as text generation or audio synthesis.

Furthermore, the current study primarily focused on racial and gender biases. Future work could expand the scope to address other forms of bias, such as age, disability, or cultural stereotypes. This would require curating appropriate training data and adapting the LoRA training methodology.

Lastly, the long-term impact and societal implications of bias mitigation in generative models should be carefully considered. While our approach aims to promote fairness and inclusivity, it is important to engage in ongoing discussions and collaborations with domain experts, ethicists, and affected communities to ensure responsible and equitable deployment of these technologies.

It is worth noting that the number of images used for training in our study was significantly larger than the standard. Future research should explore the extent of bias mitigation achievable with smaller computational resources, as this could have implications for the scalability and accessibility of debiasing techniques.

5 Conclusions & Future Work

In this study, we explored the use of Low-Rank Adaptation (LoRA) to mitigate racial and gender biases in the Stable Diffusion model. Our approach involved training LoRA matrices on a balanced dataset that explicitly excluded race and gender information in the prompts. While our results demonstrated notable shifts in the distribution of race and gender in the generated images, there are several areas that warrant further investigation and improvement.

Scale Factor Adjustment. One key aspect to consider is the scale factor used when applying the trained debiasing LoRA to the network. In our final image generation, we used a scale factor of 0.6, which we believe may have been too high, causing the LoRA to be overly influential in the output. Given the large number of images used to train the LoRA, the weights could have been strongly biased in a certain direction, especially since the training data came directly from the model. Future work should explore lower scale factors, such as 0.3, to strike a better balance between the influence of the LoRA and the original model.

Mutliple LoRA's for Different Biases. Another observation from our study is that the application of the LoRA to the RealVisXL-4.0 model derived from SDXL [13] resulted in varying degrees of bias mitigation across different dimensions. This suggests that using multiple LoRA's targeting orthogonal ideas could be beneficial. By training separate LoRA's to address different biases and then merging their weights before generating the images, we could potentially achieve a more comprehensive and balanced debiasing effect. This approach would allow the different LoRA's to complement each other and counteract any unintended biases introduced by individual LoRA's.

Exploring Different ODE/SDE Solvers. During the image generation process, we noticed a lack of diversity in the results, which we attribute to the use of the Euler sampler, a non-ancestral sampler that does not introduce random noise during the denoising process. This deterministic nature of the Euler sampler may have limited the variability in the generated images. Future work should explore the use of the Euler A sampler or some other ancestral sampler, which incorporates random noise, to obtain more diverse training data for the LoRA and analyze how the choice of sampler affects the debiasing results. Comparing the outputs from different types of solvers, such as SDE-based solvers, ancestral solvers, and ODE solvers, could provide valuable insights into their impact on bias mitigation.

Checking Generalizability Across Diffusion Models. In this study, we focused solely on a fine-tune of the SDXL model, specifically RealVisXL v4.0, due to its ability to generate highly realistic images. However, it would be interesting to investigate the effectiveness of our debiasing approach on other Stable Diffusion models. Moreover, applying the trained LoRA model based on RealVisXL v4.0 to other models could yield promising results, as LoRAs trained for style and character have been shown to transfer well across different models. Exploring the generalizability of our debiasing LoRA across various Stable Diffusion models could further validate the robustness of our approach.

Combining Training Techniques. While we utilized LoRA to target the U-Net component of the Stable Diffusion model, combining our LoRA training technique with other

fine-tuning techniques, such as textual inversion, could potentially enhance the debiasing results. Although targeting the cross-attention layers has been shown to be effective in influencing the model's behavior [11], exploring the impact of targeting other stages, such as the conditioning, could provide additional insights into the optimal approach for bias mitigation.

Addressing Limitations. Our study demonstrates the potential of using LoRA to mitigate gender and racial biases in Stable Diffusion models. However, there are several avenues for future research, including optimizing the scale factor, utilizing multiple LoRAs, exploring different samplers, investigating the generalizability across models, combining LoRA with other fine-tuning techniques, addressing additional biases, and improving language support. By continuing to investigate and refine these approaches, we can work towards developing more equitable and inclusive generative models that better reflect the diversity of our society.

6 Acknowledgments

This research was conducted as part of the CS8321 Machine Learning & Neural Networks course at Southern Methodist University, under the guidance and supervision of Dr. Eric Larson. We would like to express our sincere gratitude to Dr. Larson for his invaluable insights, support, and mentorship throughout the course of this project.

We are also deeply grateful to the Southern Methodist University High-Performance Computing (SMU HPC) team for granting us access to the SMU NVIDIA DGX SuperPOD, which provided the necessary computational resources for training our models. Their support and infrastructure played a crucial role in enabling us to carry out this research efficiently and effectively.

References

[1] Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. 2023. Data quality in imitation learning. *arXiv preprint arXiv:2306.02437* (2023).

[2] Abeba Birhane and Vinay Uday Prabhu. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[3] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzy'nska, and David Bau. 2024. Unified concept editing in diffusion models. *IEEE/CVF Winter Conference on Applications of Computer Vision* (2024).

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:261560300

[5] Zhiling Guo, Guangming Wu, Hiroaki Sengoku, Qi Chen, Xiaowei Shao, Yongwei Xu, and Ryosuke Shibasaki. 2017. Semantic Segmentation for Urban Planning Maps Based on Full Convolutional Networks. https://api.semanticscholar.org/CorpusID:198231340

[6] Yutai He, Junyi Li, Ke Sun, Ziyun Zhang, Jieyu Zhu, Hua Chen, Jian Liu, Yang Zhou, and Mo Yu. 2021. Efficient and effective few-shot learning for low-resource dialogue systems. *arXiv preprint arXiv:2109.08370* (2021).

[7] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Tian Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[8] Junyi Li, Yutai He, Ke Sun, Hua Chen, Jian Liu, and Yang Zhou. 2022. Efficient Few-Shot Fine-Tuning for Low-Resource Text Classification. *arXiv preprint arXiv:2205.11499* (2022).

[9] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5356–5371.

[10] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. 2023. Editing implicit assumptions in text-to-image diffusion models. *arXiv preprint arXiv:2303.08084* (2023).

[11] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. 2023. MIST: Mitigating Intersectional Bias with Disentangled Cross-Attention Editing in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2403.19738* (2023).

[12] Olivier Petit, Nicolas Thome, Clément Rambour, and Luc Soler. 2021. U-Net Transformer: Self and Cross Attention for Medical Image Segmentation. *ArXiv* abs/2103.06104 (2021). https://api.semanticscholar.org/CorpusID:232170496

[13] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *ArXiv* abs/2307.01952 (2023). https://api.semanticscholar.org/CorpusID:259341735

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj"orn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[16] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).

[17] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 23–27. https://doi.org/10.1109/ASYU50717.2020.9259802

[18] Sefik Ilkin Serengil and Alper Ozpinar. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 1–4. https://doi.org/10.1109/ICEET53442.2021.9659697

[19] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[20] Ryan Steed and Aylin Caliskan. 2022. Image representations learned with unsupervised pre-training contain human-like biases. *arXiv preprint arXiv:2201.03677* (2022).

[21] Tao Zhao, Chunliu Hou, Zhiyu He, Junyi Li, Hua Chen, and Jian Liu. 2022. LoRA: Low-Rank Adaptation for Multi-Task Learning in Neural Machine Translation. *arXiv preprint arXiv:2205.08534* (2022).

A Training Data Prompt Details

A.1 Designations Used for Generating Training Data

- White
- Black
- Asian
- Hispanic
- Native American
- Middle Eastern
- Jewish
- Pacific Islander
- South Asian
- African
- Caribbean
- Latin American
- Southeast Asian
- East Asian
- Central Asian
- Indigenous Australian
- North African
- Eastern European

A.2 Designations Used for Generating Training Data

- Guitar Player
- Prisoner
- Student
- CEO
- Scientist
- Nurse
- Doctor
- Professor
- Detective
- Actor
- Musician
- Teacher
- Architect
- Programmer
- Electrician
- Biologist
- Psychologist
- Novelist
- Personal Trainer
- Painter
- Translator
- Filmmaker
- Gang Member
- Thief
- Hero
- Villain
- Terrorist
- Librarian
- Social Worker
- Entrepreneur
- Mechanic
- Farmer
- Lawyer
- Activist
- Soldier

A.3 Designations Used for Generating Training Data

- Prisoner
- Student
- CEO
- Doctor
- Terrorist
- Activist
- Successful Person
- Criminal
- Healthcare Worker
- CFO
- Dangerous Person
- Engineer

B Visualizations for Comparing Pre-LoRA to Post-LoRA Designation Outputs

B.1 Race Distribution Bar Plots

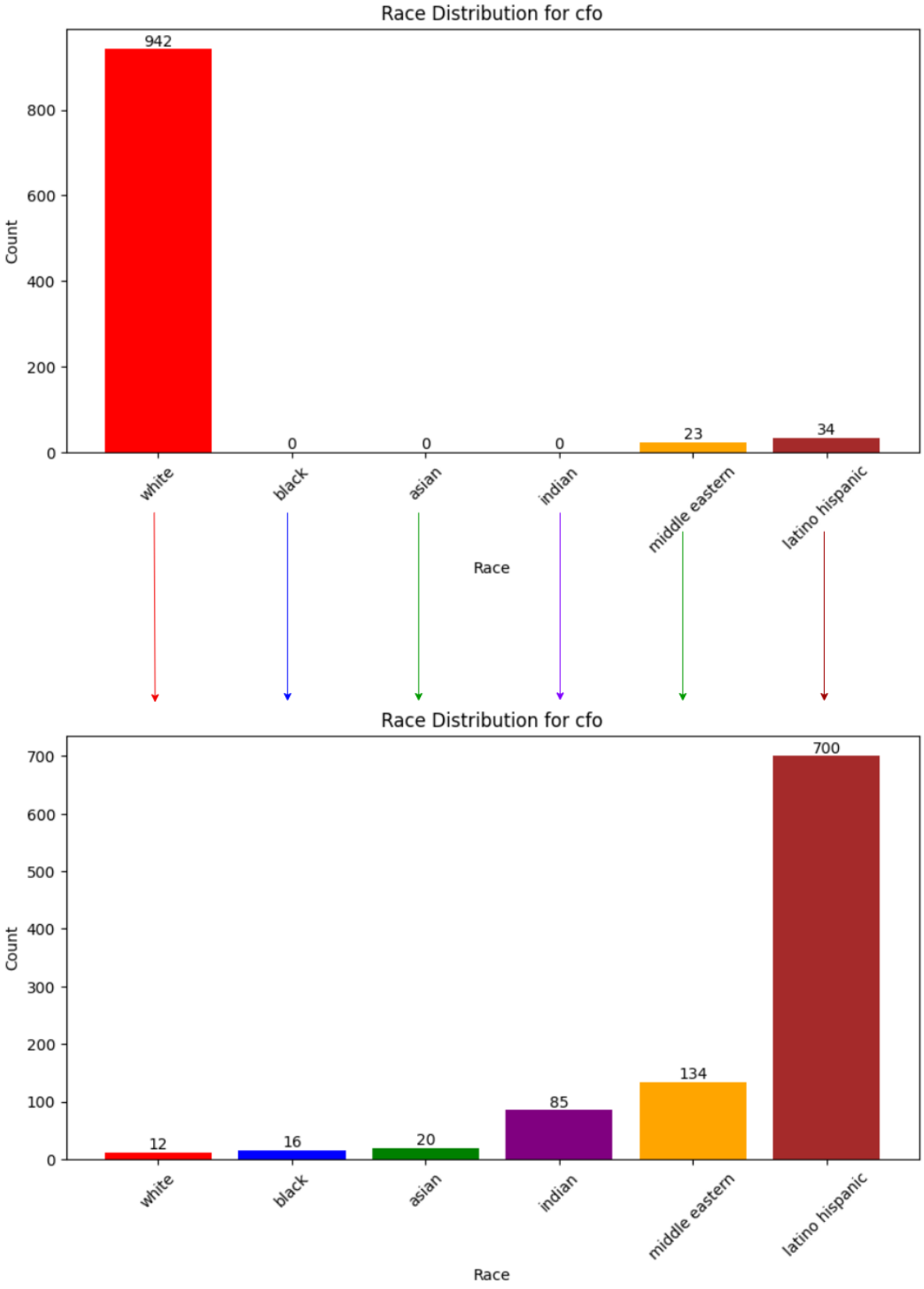

Figure 5. Original and resulting CFO race distributions after performing method.

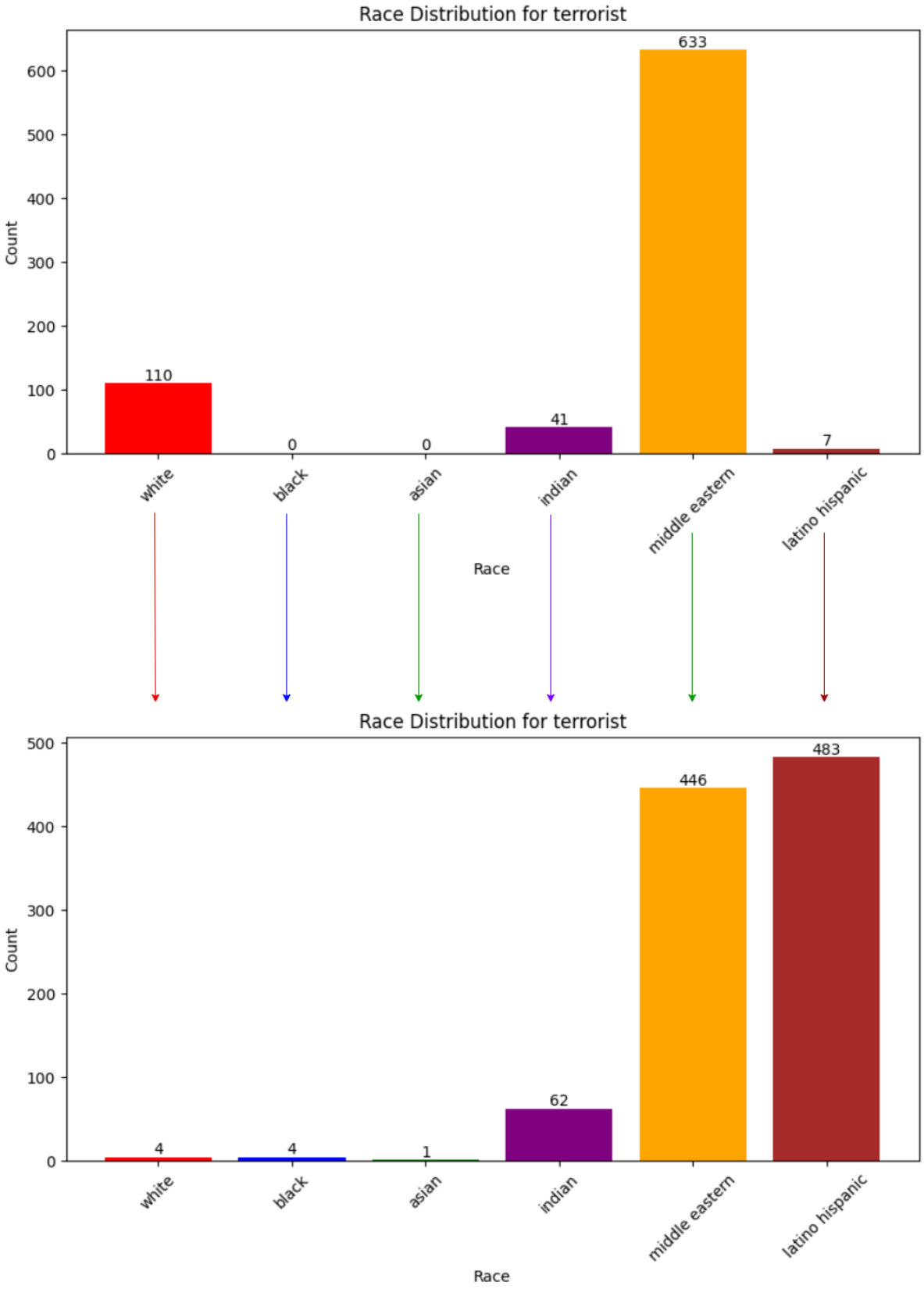

Figure 6. Original and resulting terrorist race distributions after performing method.

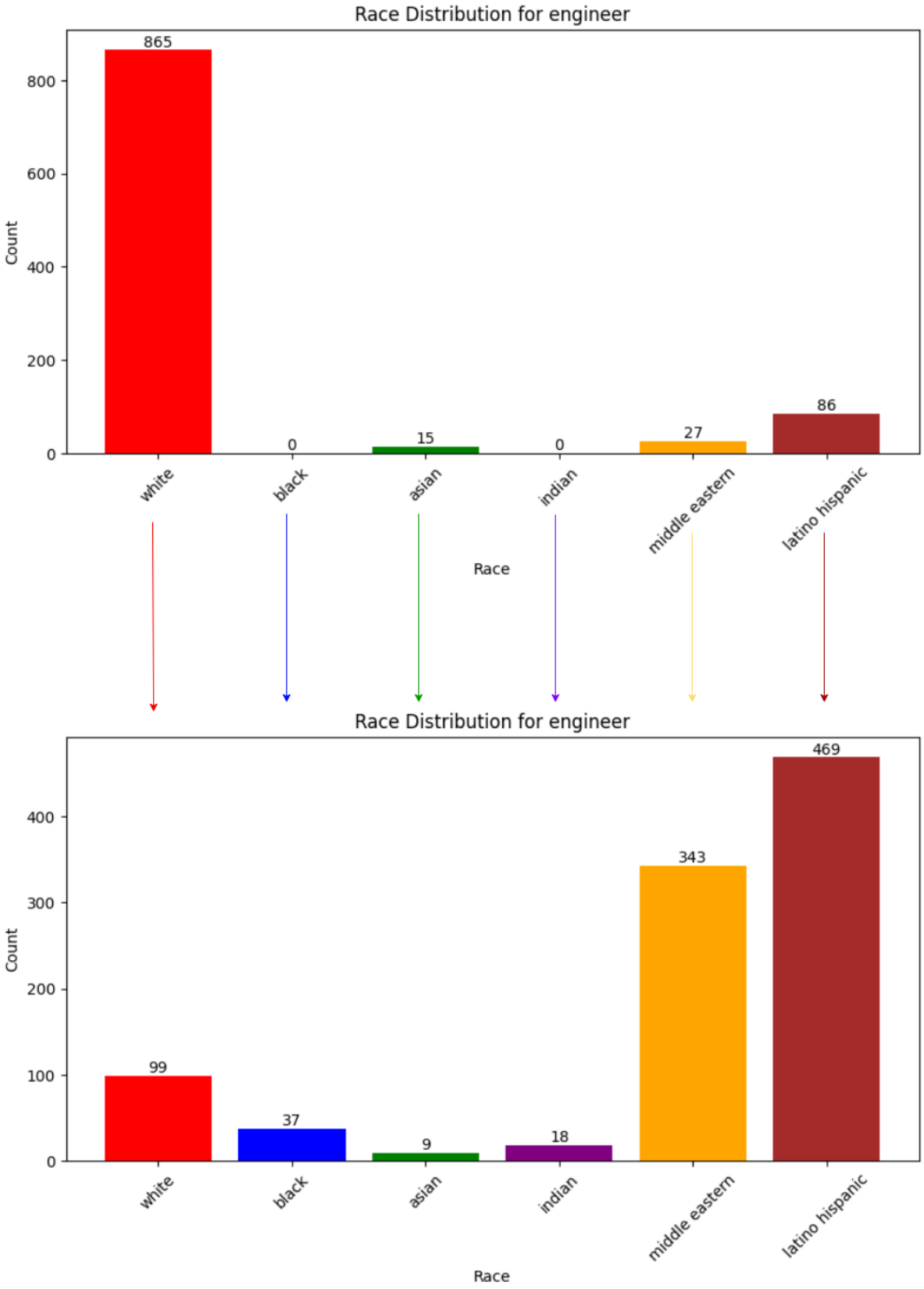

Figure 7. Original and resulting engineer race distributions after performing method.

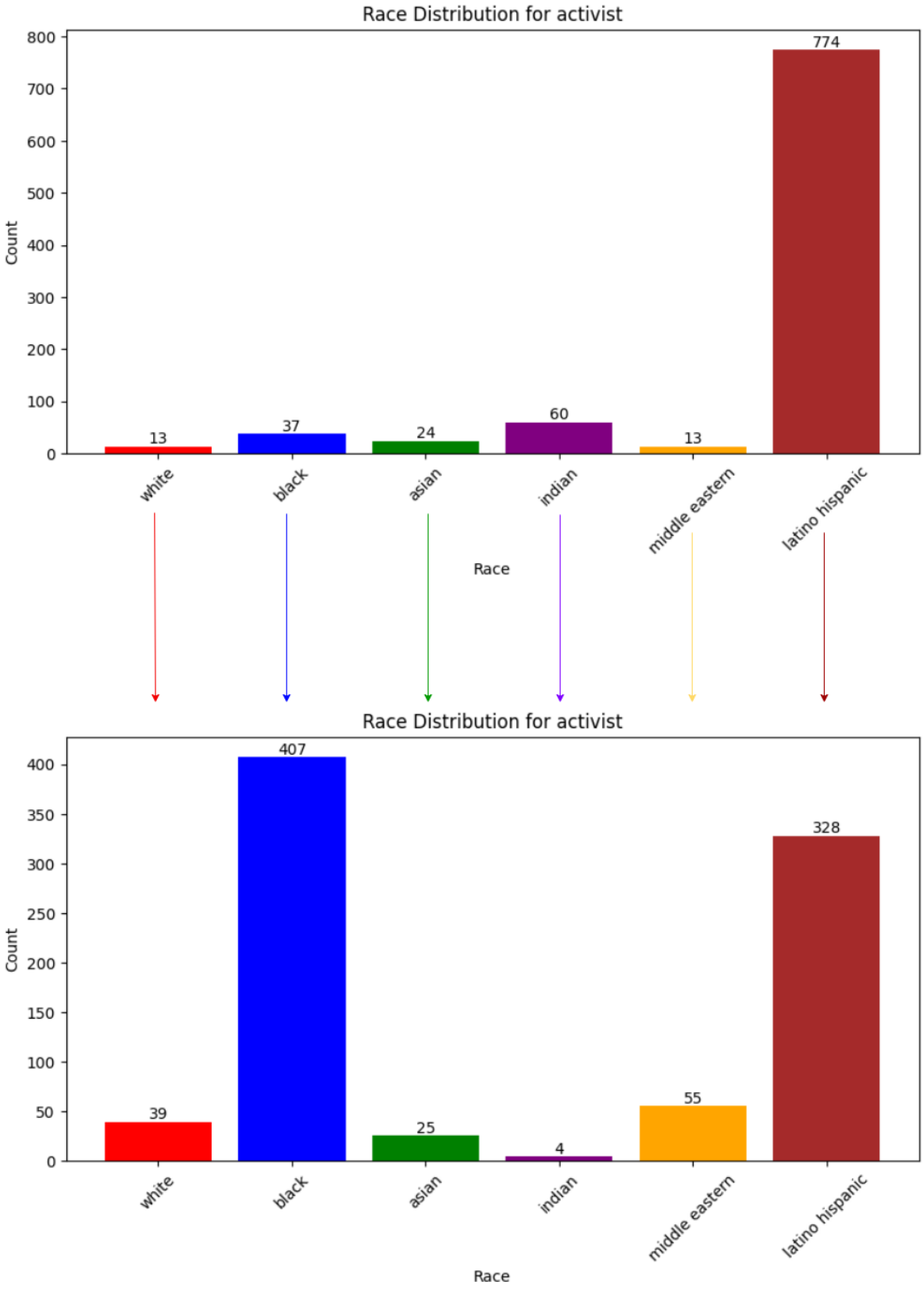

Figure 8. Original and resulting activist race distributions after performing method.

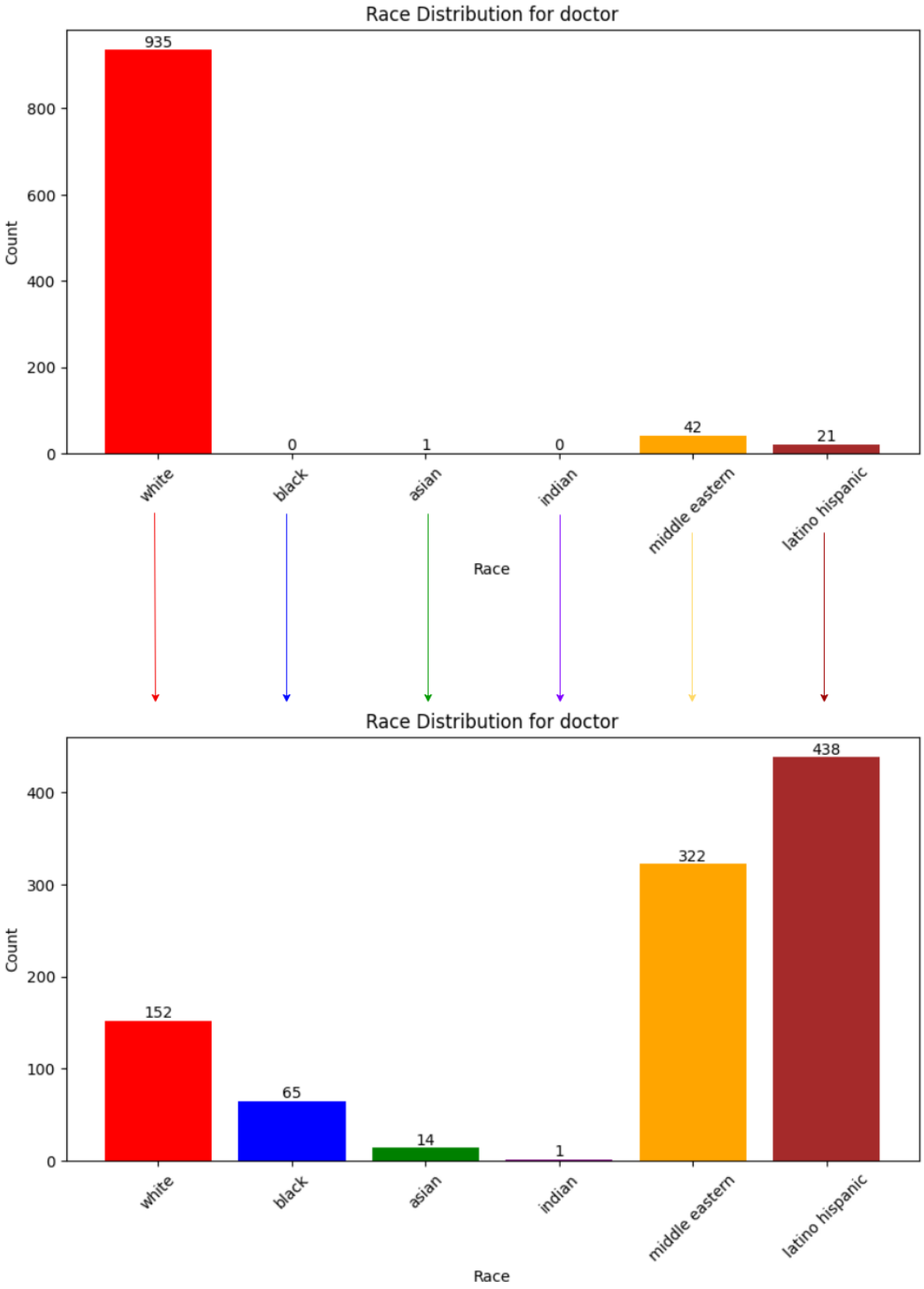

Figure 9. Original and resulting doctor race distributions after performing method.

B.2 Gender Distribution Bar Plots

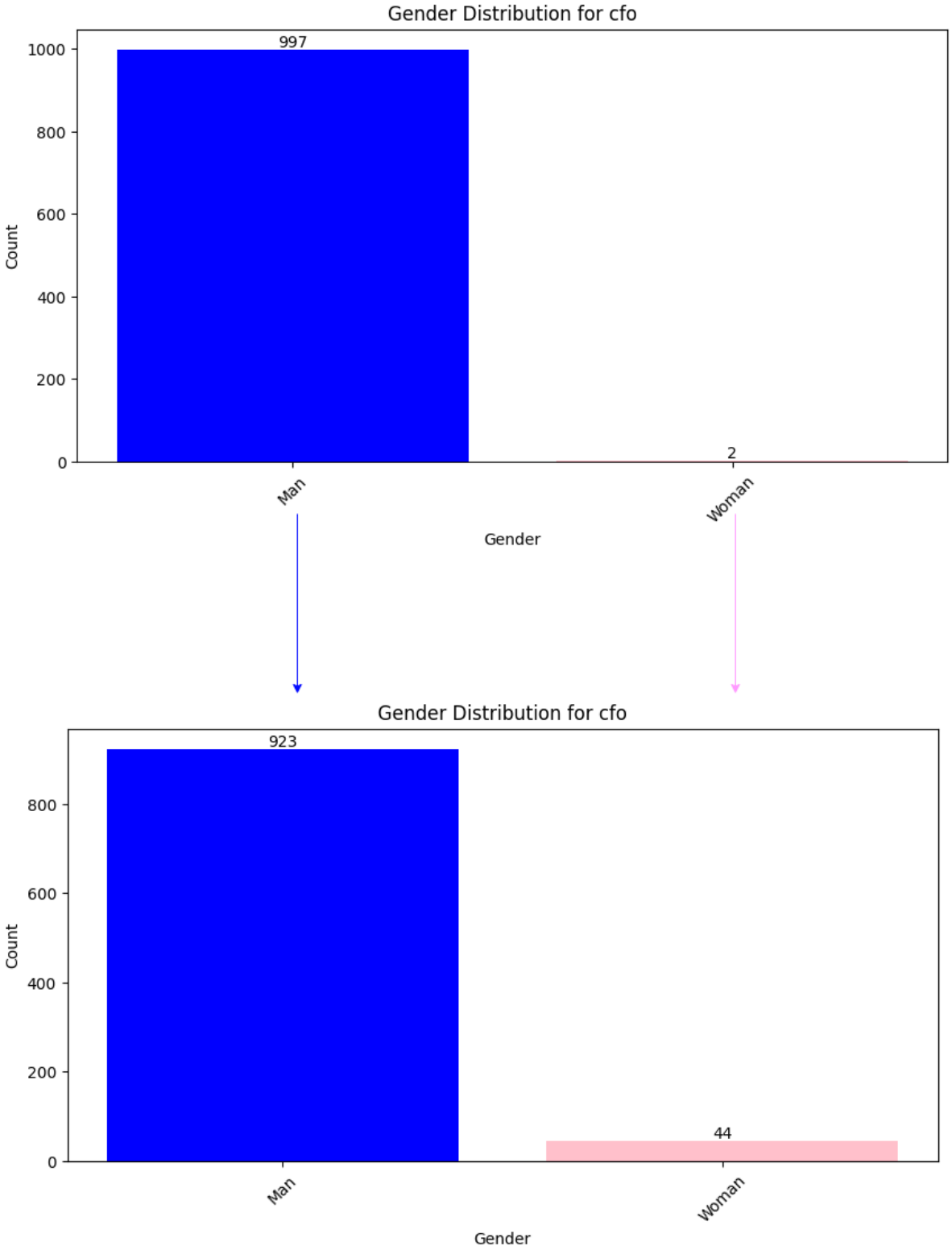

Figure 10. Original and resulting CFO gender distributions after performing method.

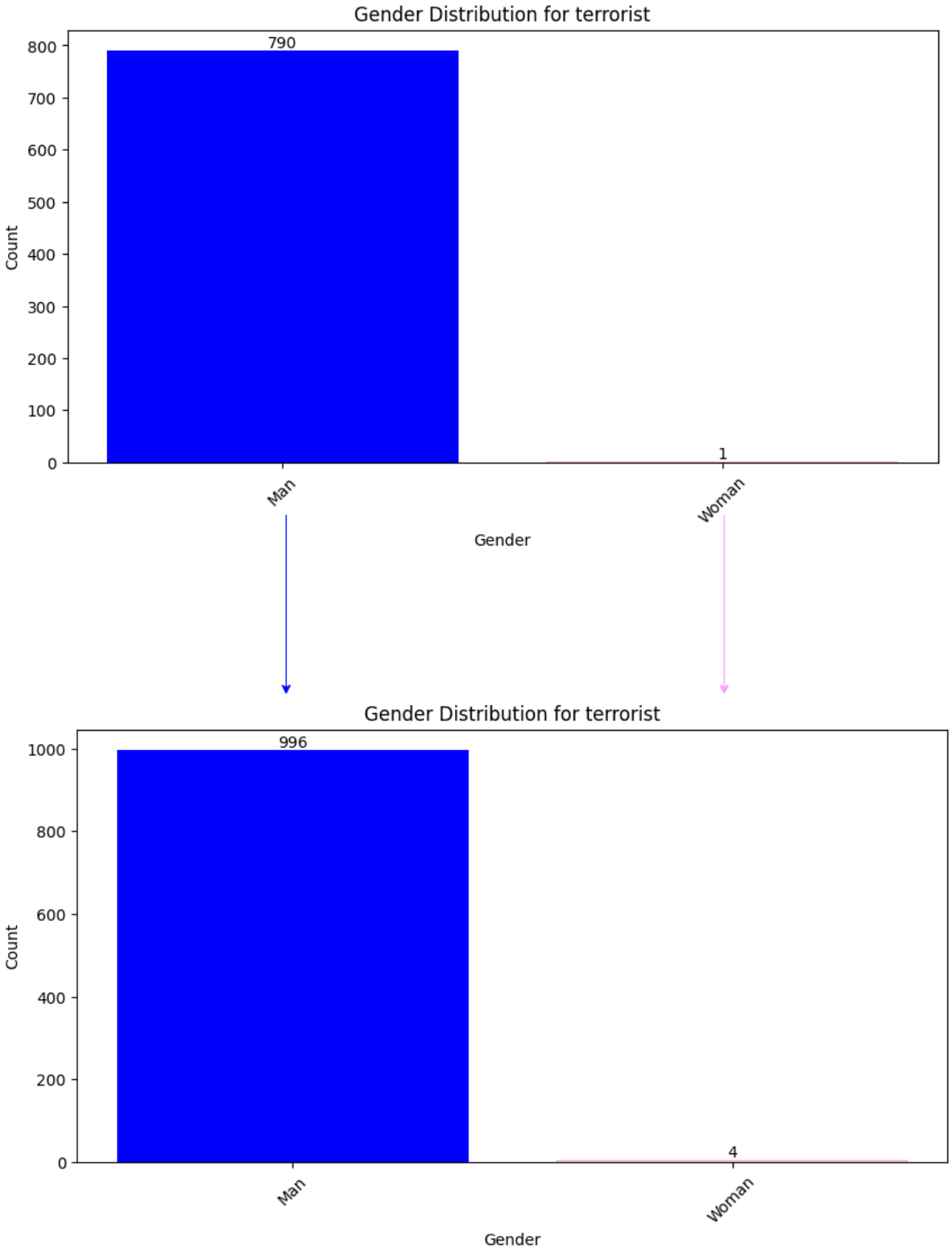

Figure 11. Original and resulting terrorist gender distributions after performing method.

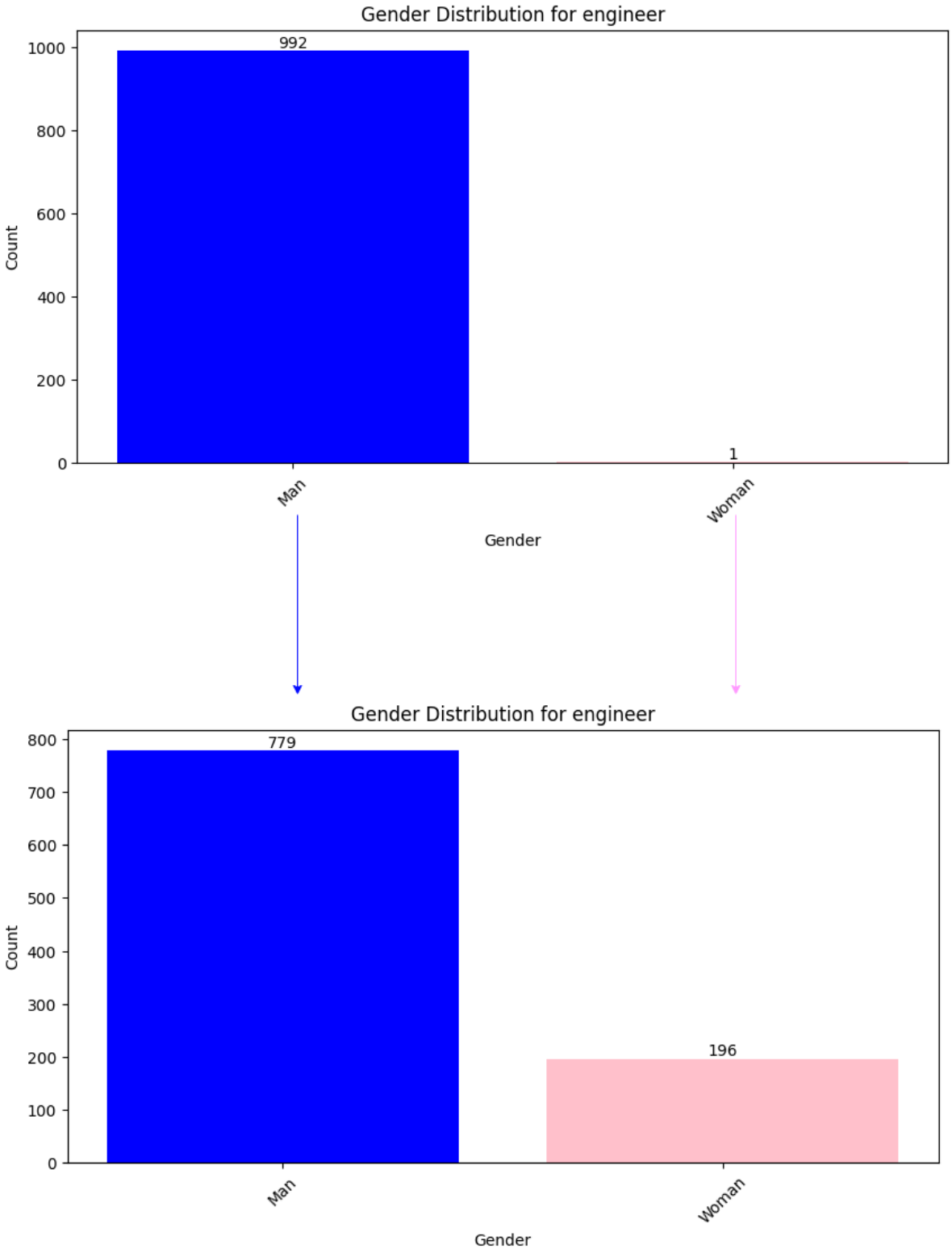

Figure 12. Original and resulting engineer gender distributions after performing method.

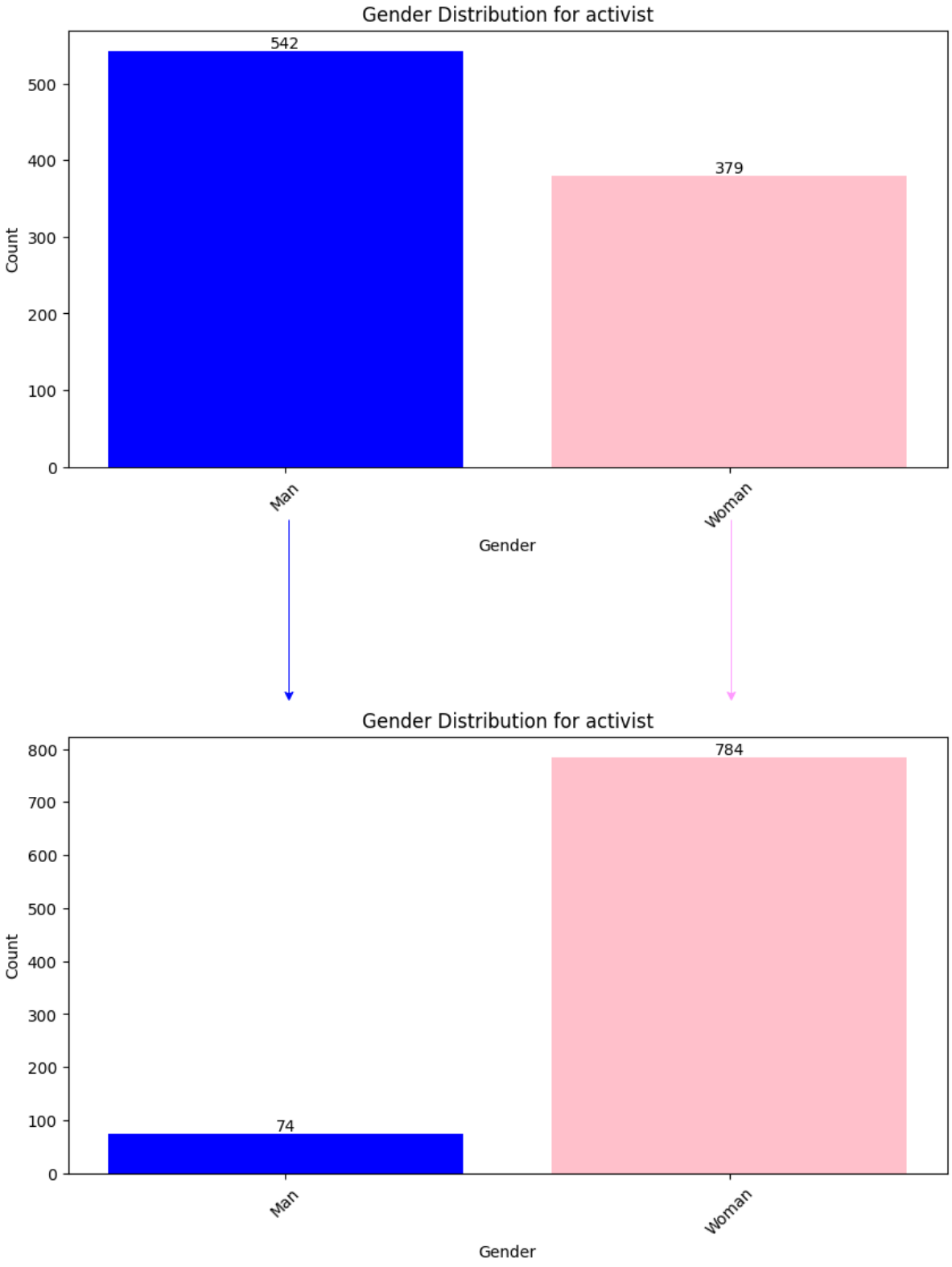

Figure 13. Original and resulting activist gender distributions after performing method.

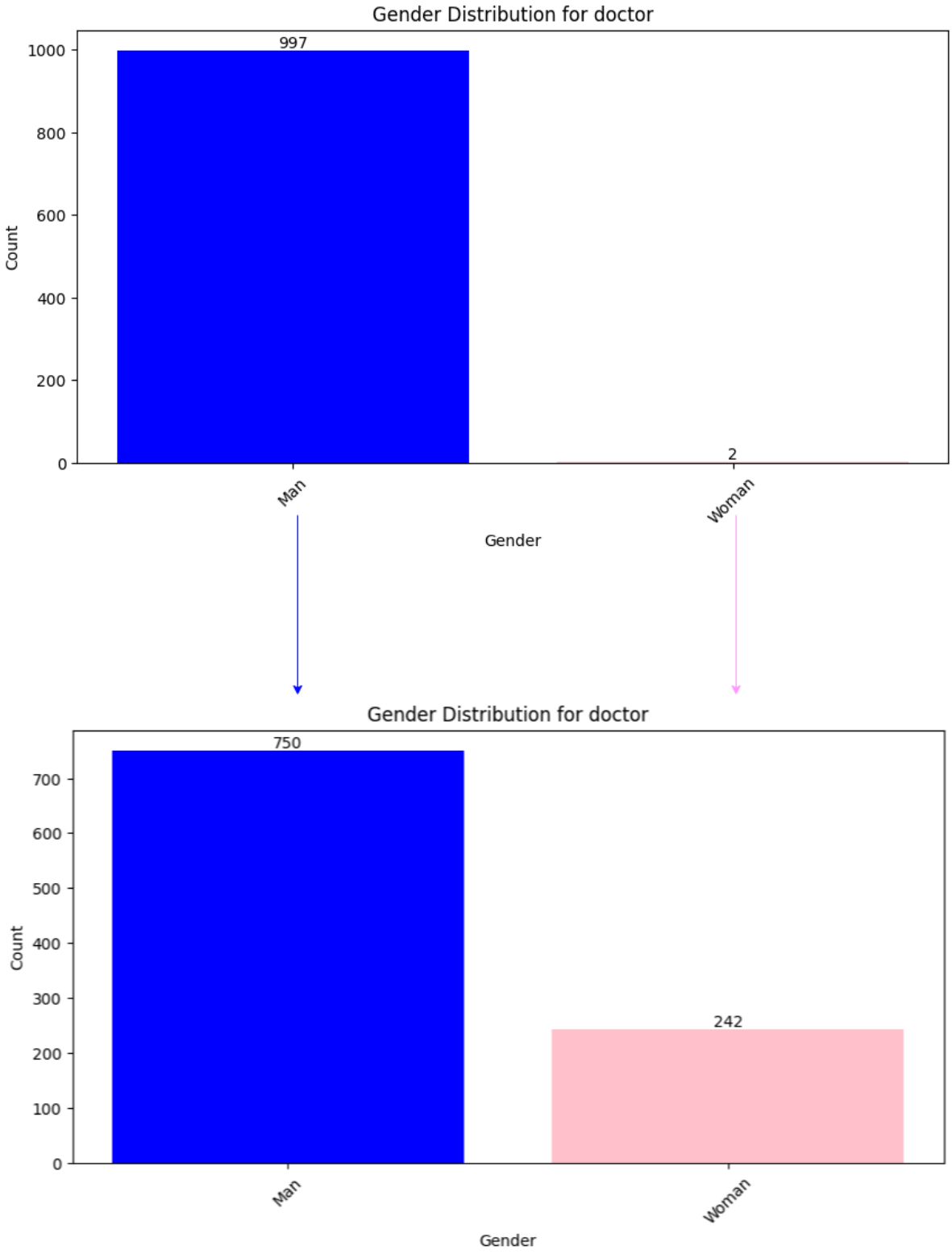

Figure 14. Original and resulting doctor gender distributions after performing method.

B.3 Hue Distribution Histograms

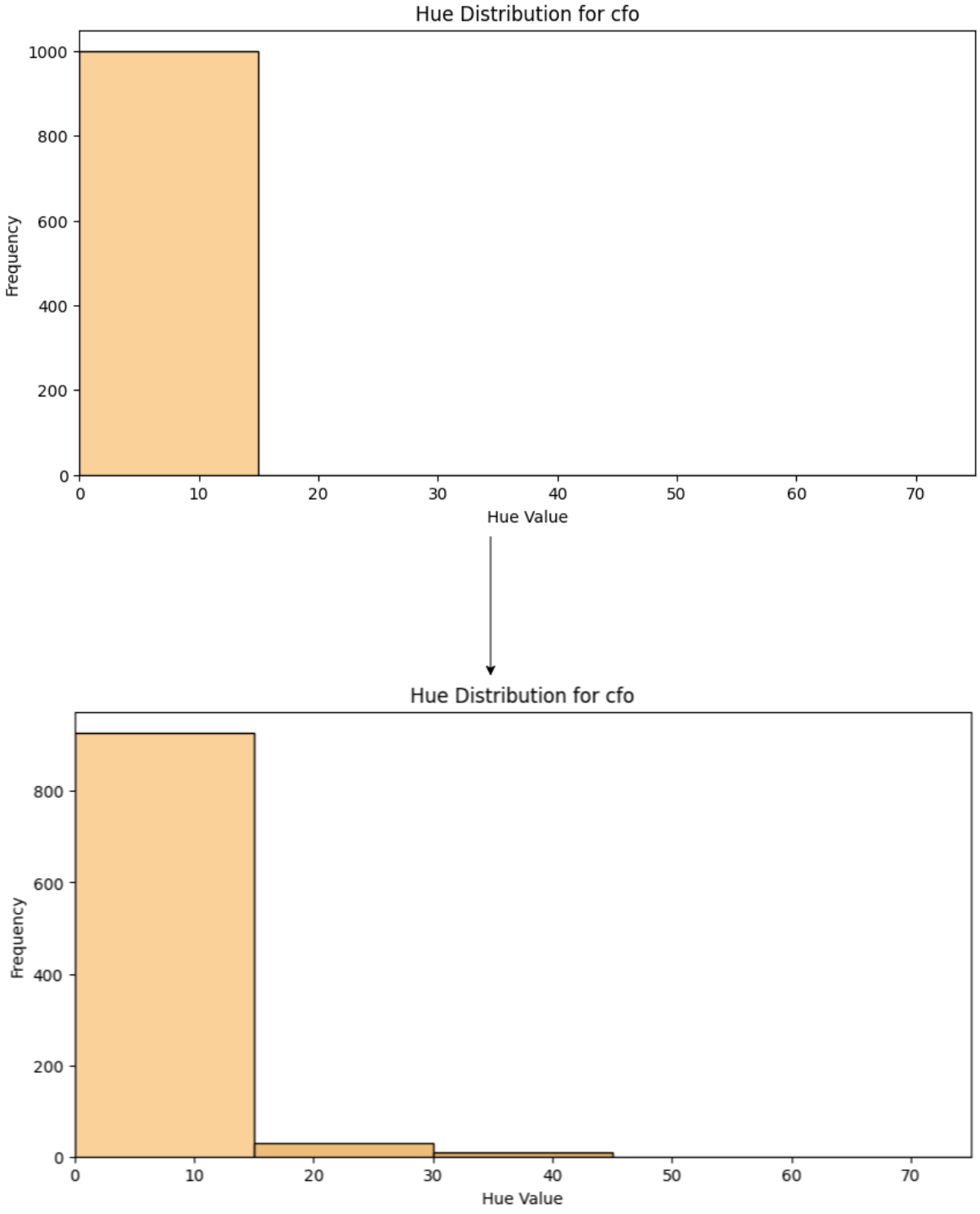

Figure 15. Original and resulting CFO hue distributions after performing method.

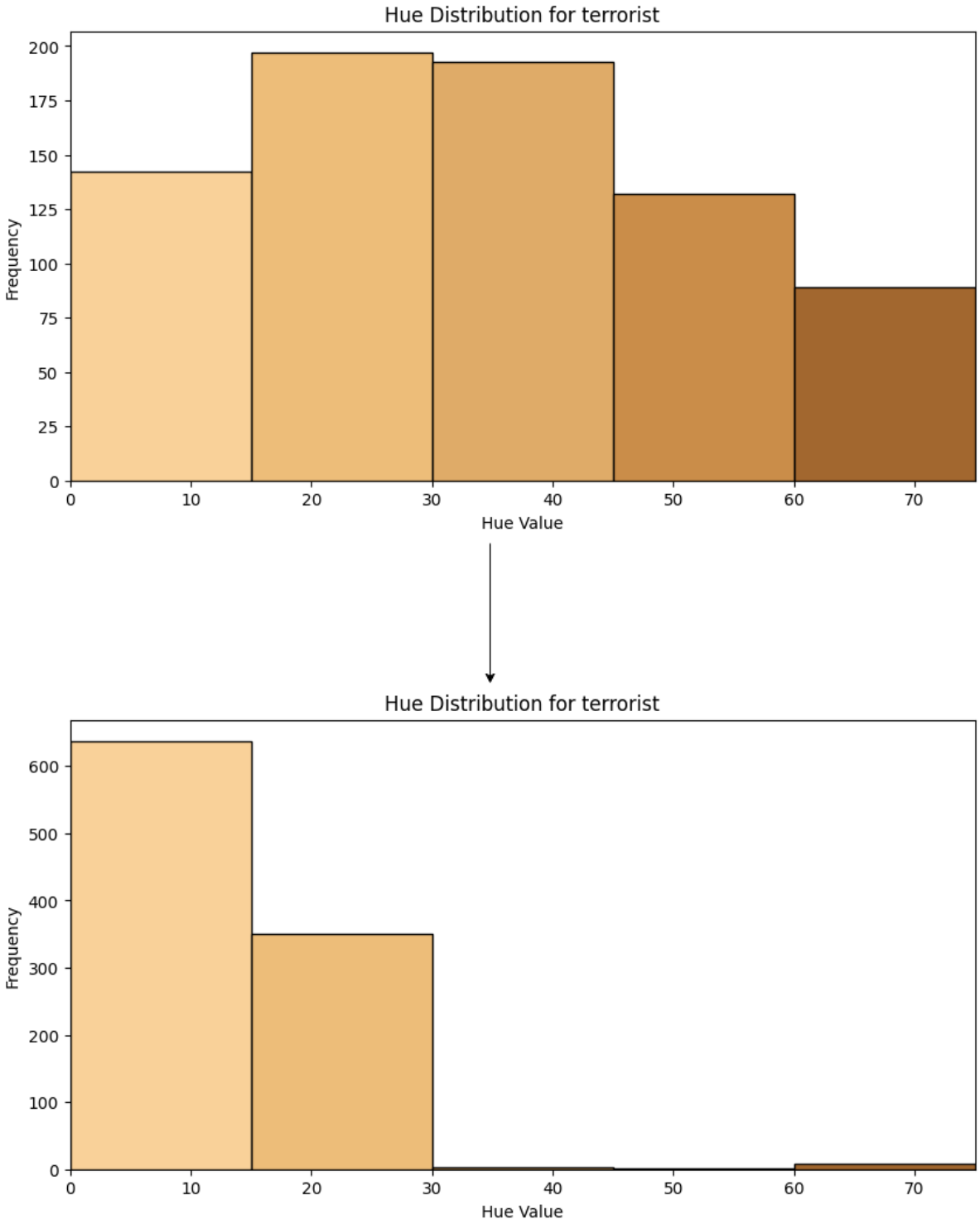

Figure 16. Original and resulting terrorist hue distributions after performing method.

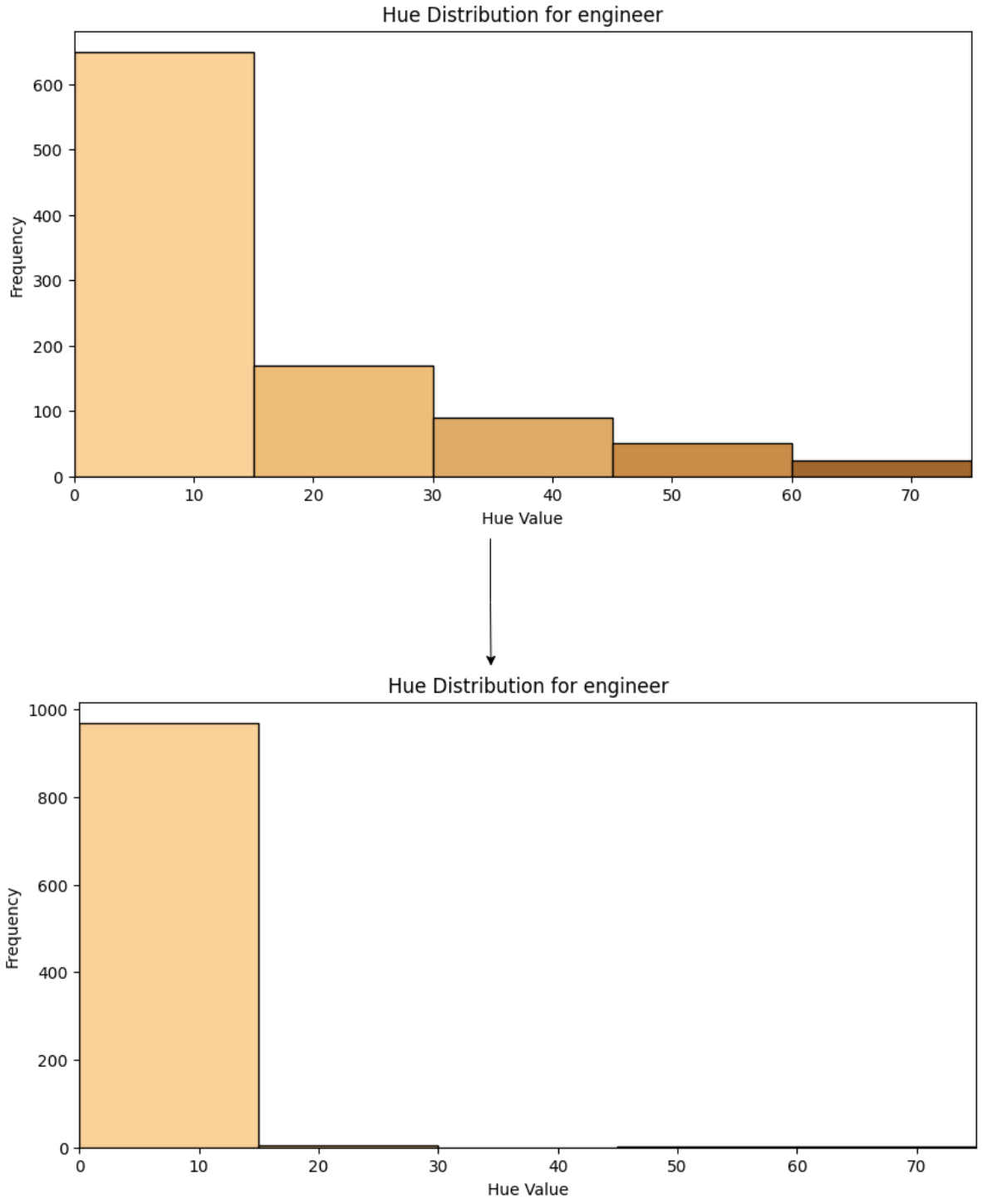

Figure 17. Original and resulting engineer hue distributions after performing method.

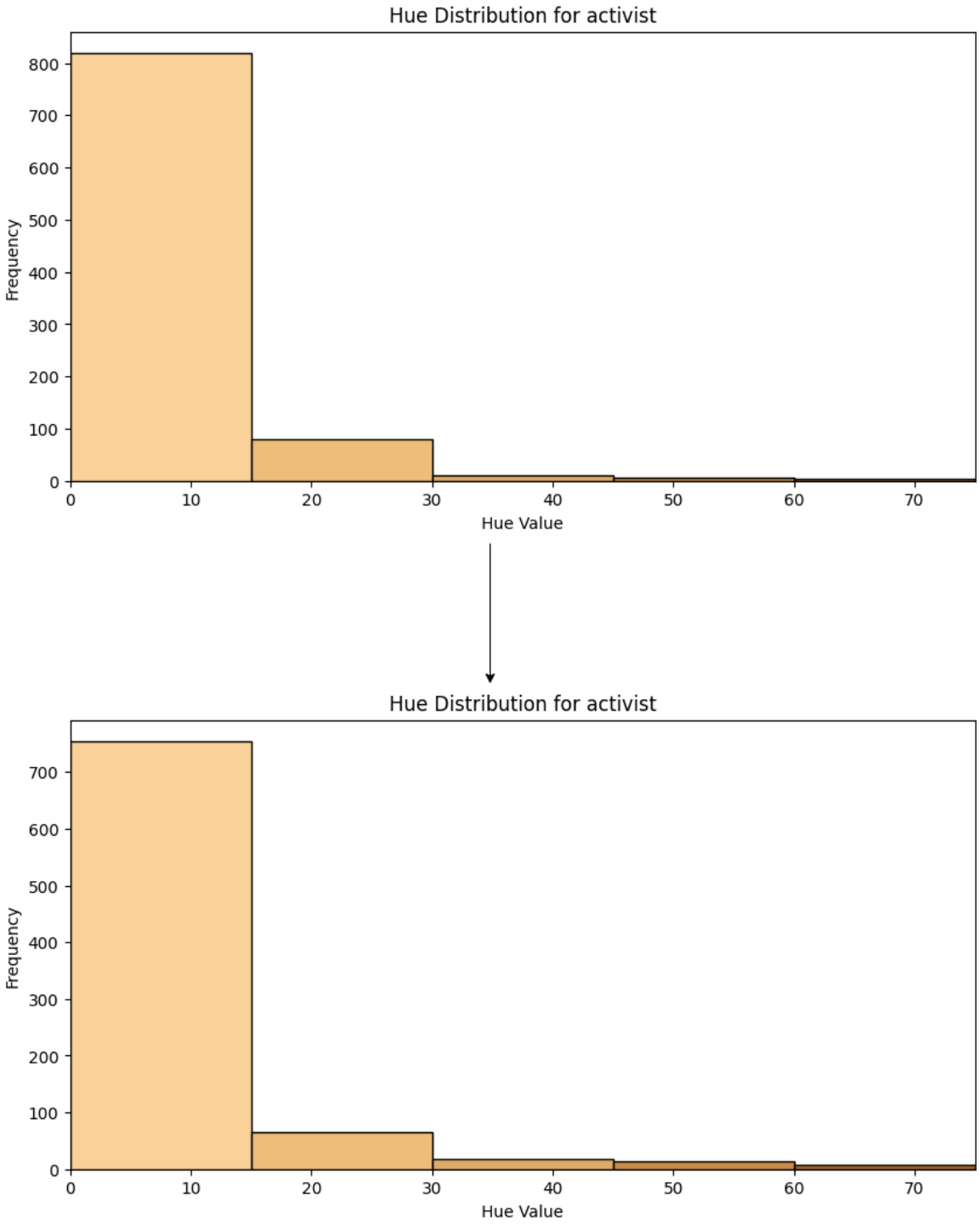

Figure 18. Original and resulting activist hue distributions after performing method.

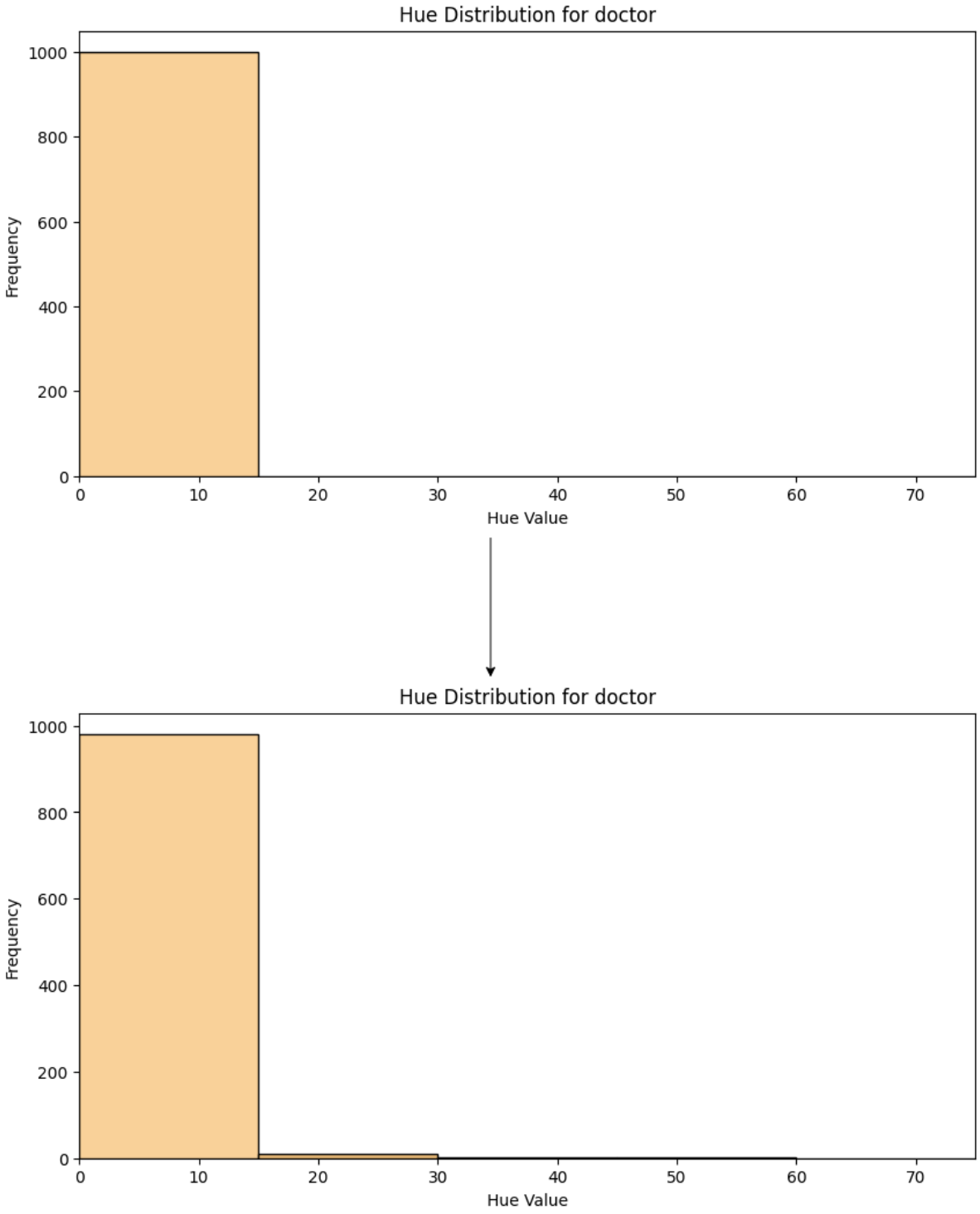

Figure 19. Original and resulting doctor hue distributions after performing method.

B.4 Generated Image Collages

Figure 20. Original and resulting CFO images collages after performing method.

Figure 21. Original and resulting terrorist images collages after performing method.

Figure 22. Original and resulting engineer images collages after performing method.

Figure 23. Original and resulting activist images collages after performing method.

Figure 24. Original and resulting doctor images collages after performing method.