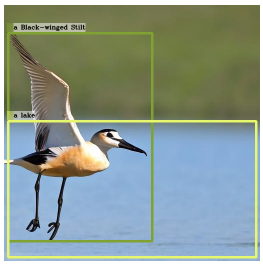
Latent Directions: A Simple Pathway to Bias Mitigation in Generative AI

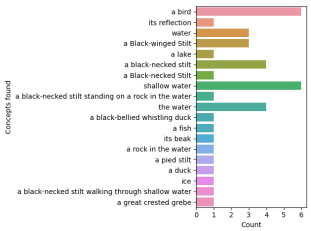
Supplementary Material

Figure 5. **Arising of the image from the noise distribution.** Noise latents saved at each diffusion step, out of 50 for the prompt "A picture of a doctor" with Stable Diffusion V2.1. Each latent is a tensor of the shape [1, 4, 64, 64]. The higher the denoising step, the more structured the final result.

6. Understanding Tool in Sec. 2



(a) Example of detections after applying O-DIG using Kosmos 2.



(b) Bar plot of detected attributes in O-DIG Kosmos 2.

Figure 6. **Frequency count of visual components with Kosmos 2 for "A picture of a waterbird" images.** To the left an example of the bounding boxes detected across generations is shown. To the right, the graph displays the count of components processed by Kosmos 2 [18] across 20 images.

Figure 7. **Kosmos 2 [18] detections in a subset of "firefighter" images.** Bounding boxes are obtained when images are run through the pipeline.

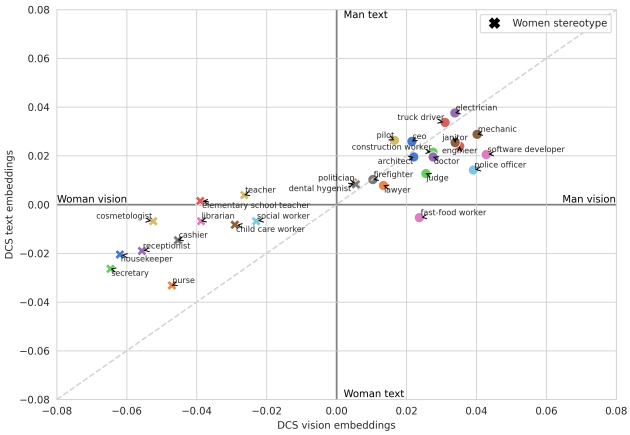


Figure 8. **Profession-Gender similarities in embeddings across CLIP's [19] text and vision encoders.** DCS stands for Difference in Cosine Similarity $DCS = CS(man/profession) - CS(woman/profession)$. When the embedding similarity is the same across encoders, the professions lie in the diagonal line.

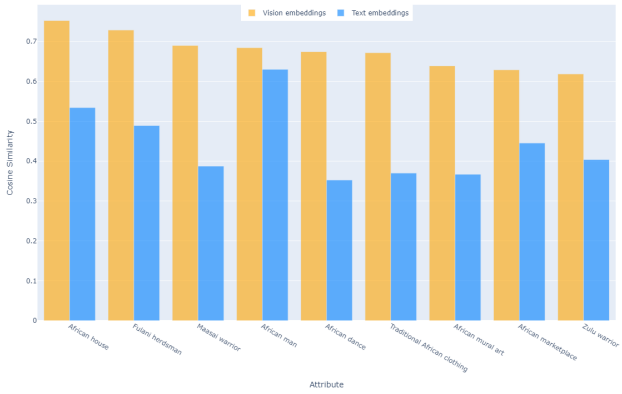


Figure 9. **Top 9 highest cosine similarities in text and vision encoders** between attributes and the concept of *A wealthy African man and his house*.

7. Mitigation[Sec. 4]

Figure 10. **Training dataset used for latent direction d_Z training.** The figure presents the 50 generated images $\tilde{x}$, with the prompt *"a photo of a man, in color, realistic, 8k"*, whose latents have been used for training.

Figure 11. **Training dataset used for latent direction d_Z training.** The figure presents the 50 generated images $\tilde{x}$, with the prompt *"a photo of a woman, in color, realistic, 8k"*, whose latents have been used for training.

Figure 12. **Training dataset used for latent direction d_Z training.** The figure presents the 50 generated images $\tilde{x}$, with the prompt *"a photo of a black man, in color, realistic, 8k"*, whose latents have been used for training.

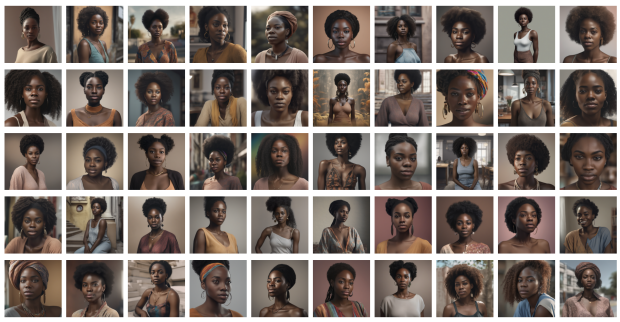


Figure 13. **Training dataset used for latent direction d_Z training.** The figure presents the 50 generated images $\tilde{x}$, with the prompt *"a photo of a black woman, in color, realistic, 8k"*, whose latents have been used for training.

Figure 14. **A hundred generated images of "a photo of a man, in color, realistic, 8k" before applying our method.** Original generations from Stable Diffusion XL [23] with an 8% of CLIP classified dark-skinned men.

Figure 16. **A hundred generated images of "a photo of a woman, in color, realistic, 8k" before applying our method.** Original generations from Stable Diffusion XL [23] with an 1% of CLIP classified dark-skinned women.

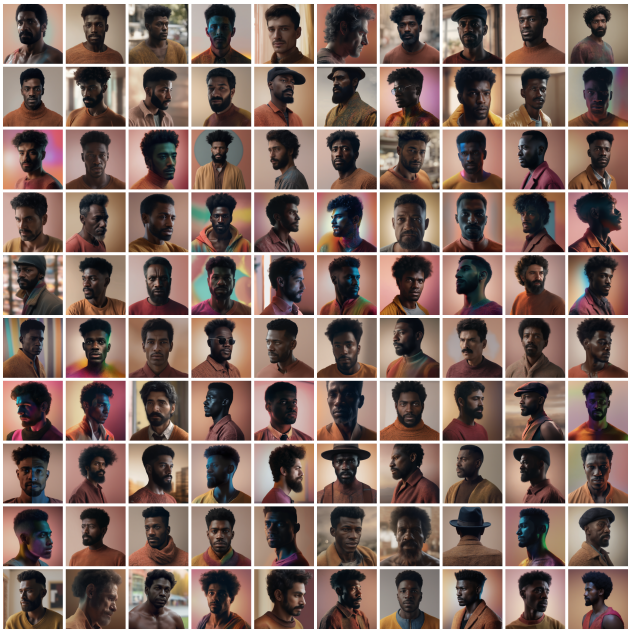


Figure 15. **A hundred generated images of "a photo of a man, in color, realistic, 8k" after applying our method with the learned *dark-skin* latent direction.** Successful transition of skin tone towards dark-skinned males with an accuracy of 95%.

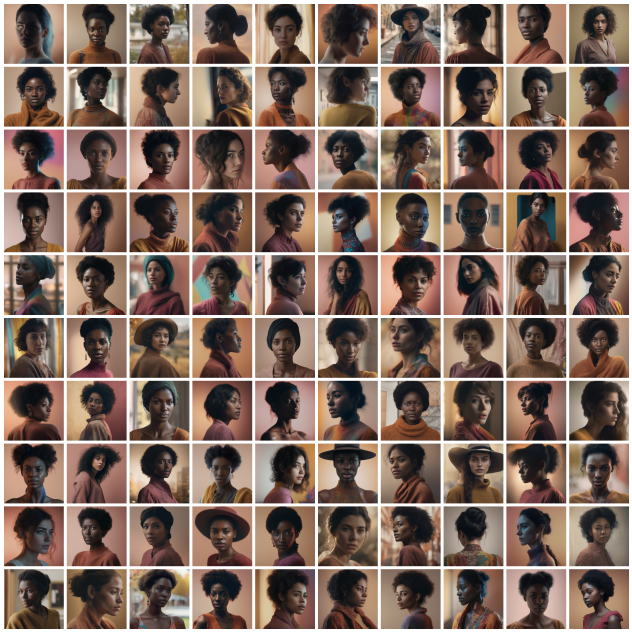


Figure 17. **A hundred generated images of "a photo of a woman, in color, realistic, 8k" after applying our method with the learned *dark-skin* latent direction.** Successful transition of skin tone towards dark-skinned females with an accuracy of 79%.

Figure 18. **A hundred generated images of "a photo of a doctor, in color, realistic, 8k" before applying our method.** Original generations from Stable Diffusion XL [23] with an 0% of CLIP classified women.

Figure 19. **A hundred generated images of "a photo of a doctor, in color, realistic, 8k" after applying our method with the learned *dark-skin* latent direction.** Successful transition towards women doctors with an accuracy of 52%.

Figure 20. **Transition between "a photo of a doctor, in color, realistic, 8k" upon the application of different latent directions.** The top row displays the original generations, followed by iterations applying the woman direction, the dark-skin direction, and a combination of both latent directions.

Figure 21. **A hundred generated images of "a photo of a firefighter, in color, realistic, 8k" before applying our method.** Original generations from Stable Diffusion XL [23] with an 14% of CLIP classified women.

Figure 22. **A hundred generated images of "a photo of a firefighter, in color, realistic, 8k" after applying our method with the learned *woman* latent direction.** A small improvement of 8% is seen after CLIP's classification.

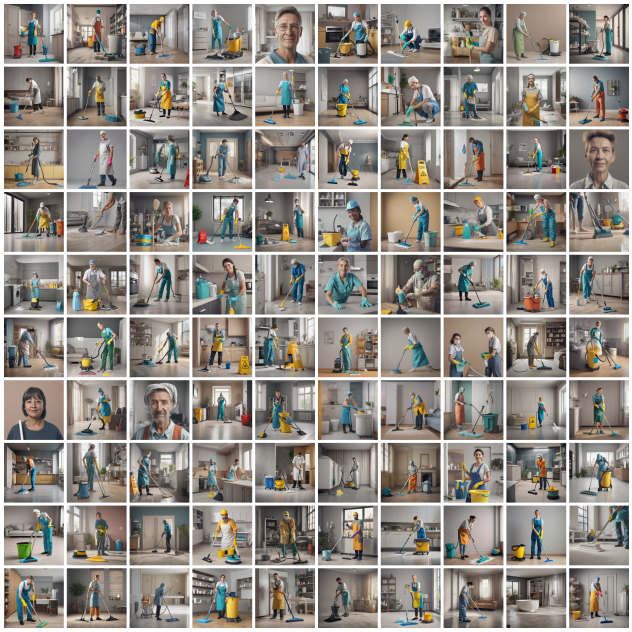


Figure 23. **A hundred generated images of "a photo of a cleaner, in color, realistic, 8k" before applying our method.** Original generations from Stable Diffusion XL [23] with an 35% of CLIP classified men.

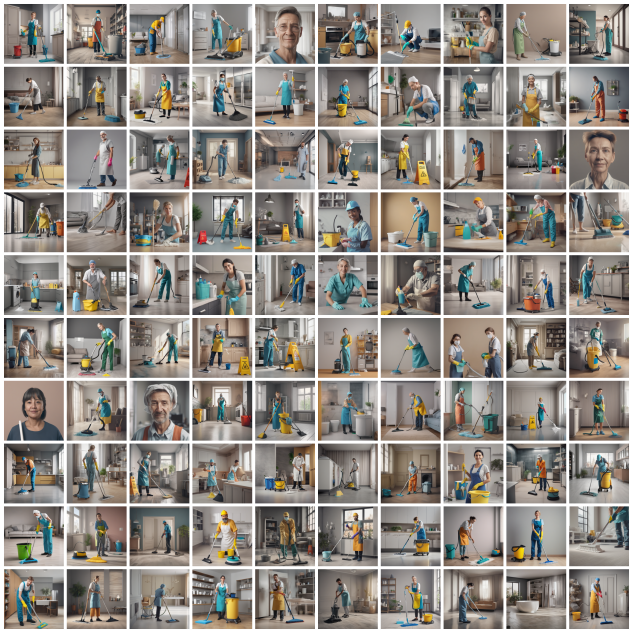


Figure 24. **A hundred generated images of "a photo of a cleaner, in color, realistic, 8k" after applying our method with the learned *man* latent direction.** A small improvement of 9% is seen after CLIP's classification.

Figure 25. **The original hundred generated images of "A picture of a waterbird".** The images show a hundred generations using Stable Diffusion V2.1 without any debiasing.

Figure 26. **A hundred generated images of "A picture of a waterbird after debiasing using the combination of our method and Chuang *et al.*'s [7].** The images show a hundred generations using Stable Diffusion V2.1 under the application of the L10 W3 latent direction configuration, in combination with the text embedding manipulation proposed by [7].

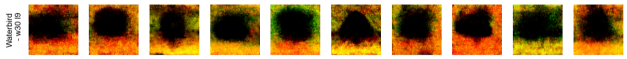


Figure 27. **Failed generation of "A picture of a waterbird" with landbird direction of weight 30.** The application of an extremely high weight brings the generation outside of the distribution, yielding poor results.

Figure 28. **Training dataset used for latent direction d_Z training.** The figure presents the 50 generated images $\tilde{x}$, with the prompt *"a photo of a wedding, in color, realistic, 8k"*, whose latents have been used for training.

Figure 29. **Training dataset used for latent direction d_Z training.** The figure presents the 50 generated images $\tilde{x}$, with the prompt *"A wealthy African man and his house."*, whose latents have been used for training.

Figure 30. **Training dataset used for latent direction d_Z training.** The figure presents the 50 generated images $\tilde{x}$, with the prompt *"a photo of an Indian wedding, in color, realistic, 8k"*, whose latents have been used for training.

Figure 31. **Training dataset used for latent direction d_Z training.** The figure presents the 50 generated images $\tilde{x}$, with the prompt *"A wealthy man and his house."*, whose latents have been used for training.

Figure 32. **A hundred generated images of "a photo of a wedding, in color, realistic, 8k" before applying our method.** All generated images exhibit a resemblance to the Western wedding concept.

Figure 33. **A hundred generated images of "a photo of a wedding, in color, realistic, 8k" after applying our method with the learned *India wedding* latent direction.** The predominance of the "red" color in them shows the association of the color with the latent direction.

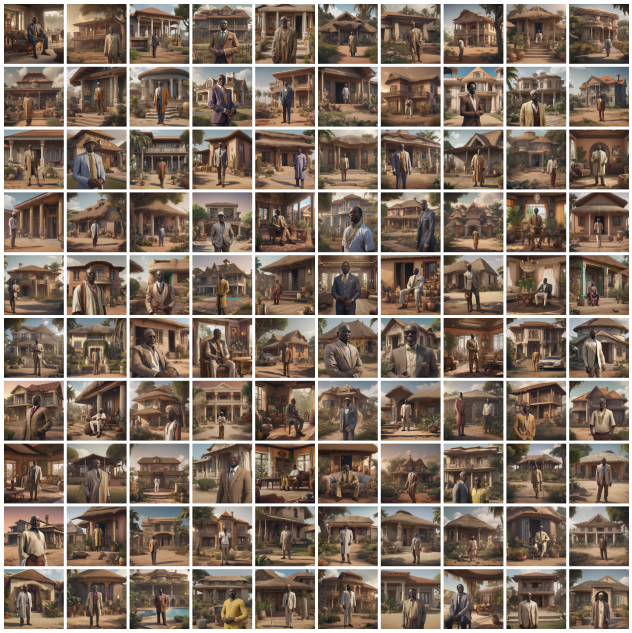


Figure 34. **A hundred generated images of "A wealthy African man and his house." before applying our method.** Most generations present houses in terracotta colors and thatched roofs.

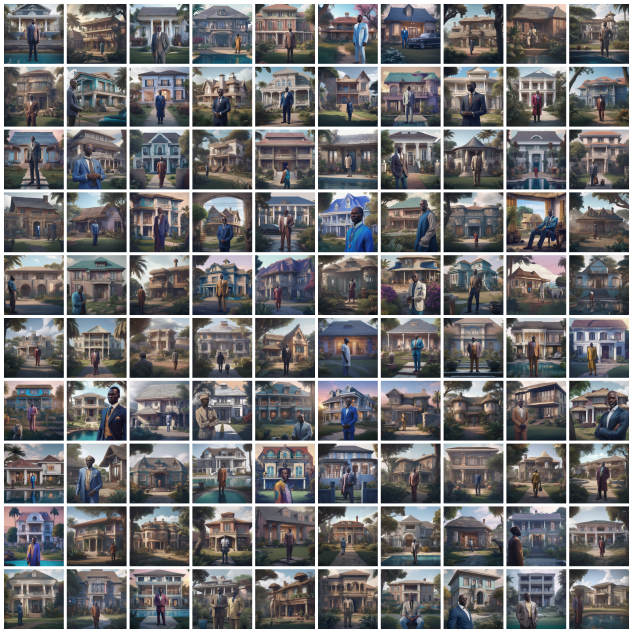


Figure 35. **A hundred generated images of "A wealthy African man and his house." after applying our method with the learned *wealthy man* latent direction.** Transition is seen from the disappearance of thatched roofs.

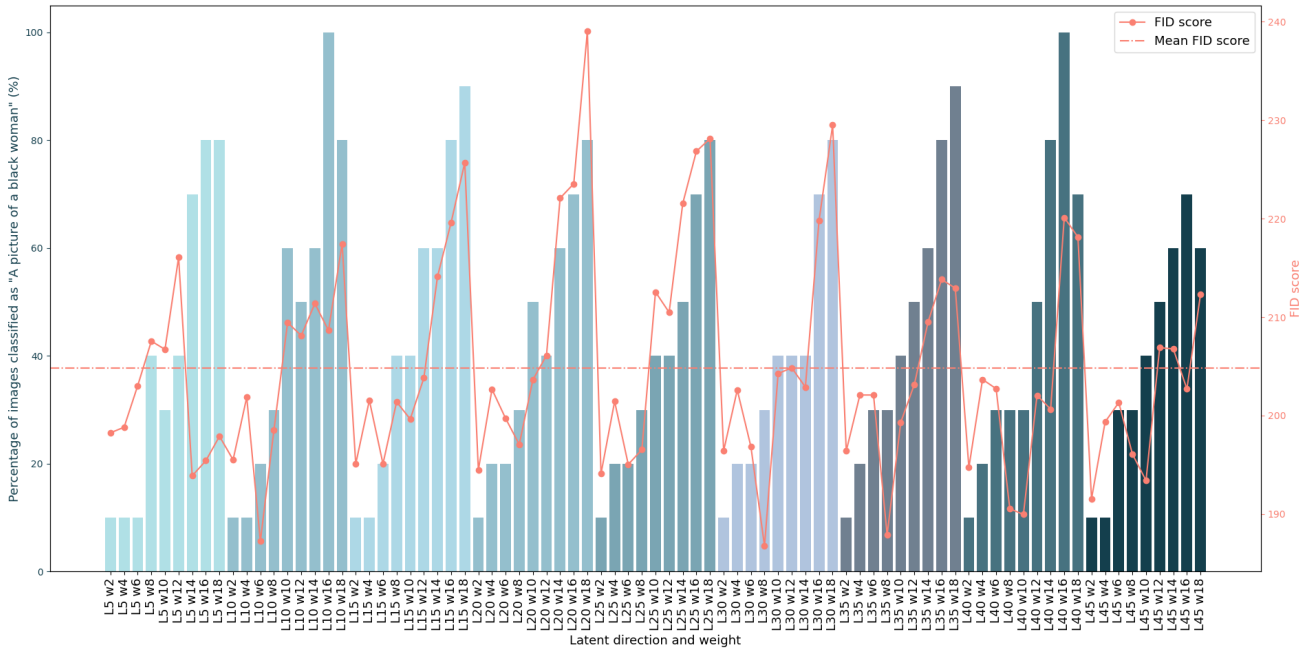


Figure 36. **Optimal configuration exploration plot.** Comparison of results at different weights and latents for "a photo of a woman, in color, realistic, 8k". The line plot presents the FID score computed between each configuration with 10 generations at the latent and weight specified, and the debiased generated dataset. Please note that the debiased dataset could also consist of real images. The bar plot presents CLIP's [19] classification results for the classes: ["A picture of a black woman", "A picture of a white woman"].

Figure 37. **Illustration of configurations identified for low and high FID scores, as well as CLIP classifications.** The top row displays the 10 generations representing the lowest FID score. The remaining rows present the configurations with a high CLIP classification for low, medium, and high FID scores.

Figure 38. **Impact of the chosen weight.** Transition at different weights from original generations to dark-skinned men. Generations obtained for the prompt "a photo of a man, in color, realistic, 8k".