BIAS MITIGATION FOR TEXT-TO-IMAGE

FairImagen: Post-Processing for Bias Mitigation in Text-to-Image Models

Zihao Fu[1], Ryan Brown[2], Shun Shao[3], Kai Rawal[2], Eoin Delaney[4], Chris Russell[2]

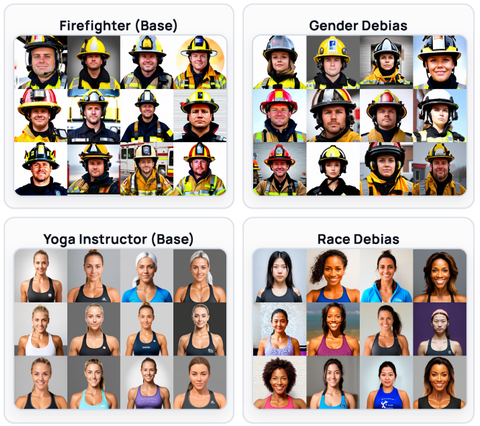
[1]The Chinese University of Hong Kong [2]University of Oxford [3]University of Cambridge [4]Trinity College Dublin

NEURAL INFORMATION PROCESSING SYSTEMS

github.com/fuzihaofzh/FairImagen

INTRODUCTION

Text-to-image models often replicate societal biases, generating stereotypical representations. FairImagen is a training-free framework that mitigates demographic biases while preserving visual quality and semantic fidelity.



FairImagen produces balanced demographic representations while preserving visual quality and semantic fidelity.

METHOD

1. Prompt Embedding Extraction
Encode prompts with CLIP to extract embeddings containing semantic and demographic information.

2. Fair Representation Transformer (FairPCA)
Learn orthogonal projection balancing variance preservation and bias removal:

$$\min_{P^\top P=I} -\mathrm{Tr}(P^\top \Sigma_X P) + \lambda \|BP\|_F^2$$

Parameter λ controls fairness-fidelity trade-off.

3. Empirical Noise Injection
Inject controlled noise along group-specific directions:

$$\tilde{E}_p'' = \tilde{E}_p' + \epsilon \cdot \delta \cdot \nu_g, \quad \delta \sim \mathcal{D}_g$$

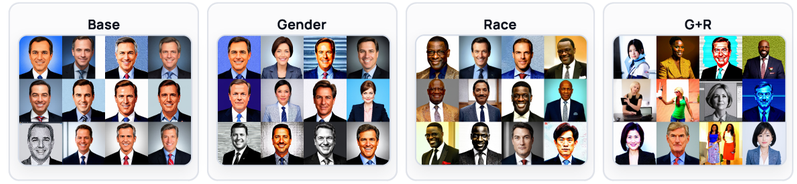
where ν_g is the bias direction and $\mathcal{D}_g$ is the empirical distribution.

4. Cross-Demographic Debiasing
Use Cartesian product of groups to debias multiple attributes jointly in a unified space.

5. Image Generator
Pass transformed embeddings to Stable Diffusion for final image generation.

CASE STUDY

We assess debiasing strategies using "a photo of a CEO" under different settings.



Base generates predominantly white males. Gender debiasing increases female representation. Race debiasing introduces racial diversity. Combined debiasing yields broader diversity.

ROBUSTNESS

We evaluate semantic alignment for prompts with contextually justified gender associations, e.g., "a middle ages blacksmith".

FairPrompt* overcorrects for historical roles, while FairImagen preserves semantic alignment and avoids overcorrection.

RESULTS

GENDER DEBIASING

METHOD	FAIR	ACC	MQ	AVG
Base	0.167	0.785	0.574	0.509
FairPrompt*	0.732	0.766	0.586	0.695
ForcePrompt	0.292	0.755	0.601	0.549
SAL	0.217	0.779	0.602	0.533
CDA	0.547	0.772	0.549	0.623
TBIE	0.350	0.782	0.567	0.566
SDID	0.507	0.776	0.553	0.612
SDID-AVG	0.315	0.783	0.562	0.553
ITI	0.270	0.769	0.528	0.522
FairImagen	0.560	0.771	0.541	0.624

RACE DEBIASING

METHOD	FAIR	ACC	MQ	AVG
Base	0.193	0.785	0.574	0.517
FairPrompt*	0.444	0.752	0.566	0.587
ForcePrompt	0.266	0.761	0.574	0.534
SAL	0.262	0.788	0.607	0.552
CDA	0.358	0.772	0.537	0.556
TBIE	0.366	0.762	0.532	0.553
SDID	0.370	0.770	0.537	0.559
SDID-AVG	0.361	0.769	0.544	0.558
ITI	0.214	0.770	0.530	0.504
FairImagen	0.389	0.760	0.536	0.562

GENDER+RACE

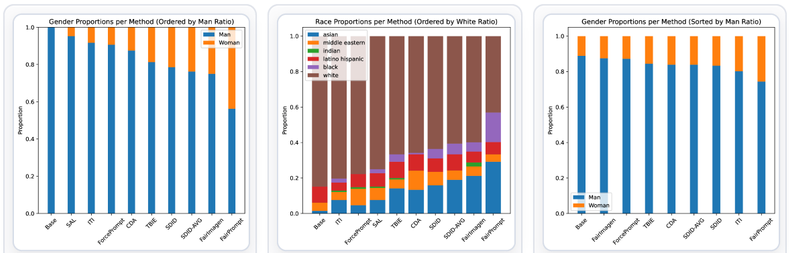
METHOD	G-F	ACC	MQ	AVG
Base	0.163	0.785	0.574	0.508
FairPrompt*	0.690	0.747	0.574	0.671
ForcePrompt	0.287	0.764	0.591	0.547
SAL	0.182	0.776	0.599	0.519
CDA	0.362	0.779	0.557	0.566
TBIE	0.400	0.776	0.546	0.574
SDID	0.223	0.782	0.556	0.520
SDID-AVG	0.352	0.778	0.553	0.561
ITI	0.320	0.747	0.467	0.511
FairImagen	0.537	0.753	0.544	0.611

Key Findings: (1) FairImagen beats general post-hoc baselines across gender, race, and joint debiasing. (2) Achieves best balance among fairness, accuracy, and image quality. (3) FairPrompt* tops scores but needs manual prompts. (4) FairImagen offers a practical, automated alternative.

* Oracle baseline requiring manual prompt engineering.

OCCUPATION ANALYSIS

We analyze occupation prompts with strong demographic biases. FairImagen significantly improves gender and race balance in biased occupations.



Male-dominated occupations | White-dominated occupations | Historical context

Key findings: FairImagen substantially reduces demographic overrepresentation in biased occupations while preserving justified associations in historical contexts.

CONTRIBUTIONS

1. Post-hoc fairness framework: Integrate FairPCA into diffusion without retraining
2. Empirical noise injection: Inject noise to improve fairness-performance trade-offs
3. Cross-demographic debiasing: Debias multiple attributes jointly without over-pruning
4. Extensive evaluation: Outperform baselines in gender, race, and joint debiasing

REFERENCES

[Friedrich et al. 2023] Felix Friedrich, Manuel Brack, Lukas Struppek, et al. "Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness." arXiv:2302.10893, 2023.
[Shao et al. 2023] Shun Shao, Yftah Ziser, and Shay B. Cohen. "Gold Doesn't Always Glitter: Spectral Removal of Linear and Nonlinear Guarded Attribute Information." EACL 2023.
[Zmigrod et al. 2019] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. "Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology." arXiv:1906.04571, 2019.
[Tanjim et al. 2024] Md Mehrab Tanjim, Krishna Kumar Singh, Kushal Kafle, et al. "Discovering and Mitigating Biases in CLIP-based Image Editing." WACV 2024.
[Li et al. 2024] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. "Self-discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation." CVPR 2024.
[Zhang et al. 2023] Cheng Zhang, Xuanbai Chen, Siqi Chai, et al. "ITI-GEN: Inclusive Text-to-Image Generation." ICCV 2023.
[Kleindessner et al. 2023] Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. "Efficient Fair PCA for Fair Representation Learning." AISTATS 2023.