FairImagen

Fair Image Generation via Post-Hoc Debiasing

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

GitHub:
github.com/fuzihaofzh/FairImagen

NeurIPS 2025

FairImagen Mitigates Demographic Biases

⚠ **The Problem: Bias in Generative Models**

Text-to-image models (Stable Diffusion, DALL·E, Imagen) often generate images that:

- Overrepresent certain demographics (e.g., white males for "CEO")
- Underrepresent minorities in professional roles
- Perpetuate harmful stereotypes and societal biases

✓ **Our Solution:** FairImagen applies post-hoc debiasing without retraining models

Gender Debiasing

Firefighter (Baseline)

Firefighter (FairImagen)

Race Debiasing

Yoga Instructor (Baseline)

Yoga Instructor (FairImagen)

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

Motivation: Bias in Text-to-Image Models

- **Text–to–image models** (Stable Diffusion, DALL·E, Imagen) produce photorealistic images from prompts
- **Problem:** They replicate and amplify societal biases
- **Examples:**
  - "a photo of a CEO" → white males
  - "a nurse" → females
- **Impact:** Concerns about fairness, representation, and downstream harms

Goal: Mitigate demographic biases while preserving image quality and semantic fidelity

Existing Debiasing Approaches

Prompt-based

Idea: Modify input prompts to influence outputs

Examples:

- Fair Diffusion: fairness–guided prompts
- FairT2I: LLM–based prompt revision
- EquiPrompt: learned fairness prompts

Issue: Heuristic rewriting, requires per–image manual effort

Fine-tuning

Idea: Update model parameters for fairness

Examples:

- Fair Mapping: linear projection layer
- ITI: align with fair visual examples
- Concept Editing: modify specific modules

Issue: Computationally intensive, requires model access

Post-hoc Editing

Idea: Modify prompt embeddings at inference

Examples:

- SDID: subtract bias directions
- TBIE: PCA–based debiasing
- ITI–GEN, SAL, CDA, FairQueue

✓ **Advantage:** Training–free, deployment–friendly, model–agnostic

Criteria	Prompt–based	Fine–tuning	Post–hoc editing
Training–free	✓	✗	✓
Black–box compatible	✓	✗	✓
Low human effort	✗	✓	✓
Low computational cost	✓	✗	✓
Generalizable to new prompts	✗	✓	✓
Strong bias mitigation	✗	✓	✓
Preserves prompt fidelity	✓	✓	✗
Easy deployment	✗	✗	✓

Our Focus: Post–hoc methods offer the best balance of effectiveness, efficiency, and deployability

FairImagen: Our Approach

Key Idea

Integrate **Fair Principal Component Analysis (FairPCA)** into Stable Diffusion pipeline to remove demographic information while preserving semantic content

Three Main Components

1. **Prompt Embedding Extractor:** Extract CLIP embeddings
2. **Fair Representation Transformer:** Apply FairPCA projection
3. **Image Generator:** Generate images from transformed embeddings

Enhancements: Empirical noise injection + unified cross-demographic debiasing

Fair Representation Transformer

Classical PCA Objective

$$\arg\min_{P^\top P=I} \sum_{i=1}^{n} \|\mathbf{x}_i - PP^\top \mathbf{x}_i\|_2^2$$

Minimizes reconstruction error when projecting to lower dimensions

FairPCA: Add Fairness Constraint

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

- First term: Preserve semantic content (reconstruction quality)
- Second term: Remove group–specific information (fairness)
- $B = Z^\top X$: Group–dependent feature matrix
- λ: Trade–off hyperparameter

Empirical Noise Injection

Motivation

FairPCA removes all demographic signals → overly neutral outputs (e.g., feminine–looking men)

Solution: Add controlled noise along bias directions to maintain diversity

Method

1. Compute bias direction for group g:

$$\nu_g = \frac{1}{|X^{(g)}|} \sum_{\bar{E}_p \in X^{(g)}} \bar{E}_p - \bar{E}$$

2. Build empirical distribution:

$$\mathcal{D}_g = \{\nu_g^\top \bar{E}_p : \bar{E}_p \in X^{(g)}\}$$

3. Inject noise during inference:

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

Cross-Demographics Debiasing

Challenge

Naively stacking multiple demographic attributes forces features to be orthogonal to each group direction → degraded visual quality

Our Solution: Unified Joint Space

Construct attribute space from **Cartesian product** of all groups

Example:

- Gender: {Male, Female}
- Race: {White, Asian, Black}

$$\mathcal{A}_{\text{joint}} = \{\text{White Male}, \text{White Female}, \text{Asian Male}, \text{Asian Female}, \text{Black Male}, \text{Black Female}\}$$

Apply FairPCA **once** over this joint space → simultaneous debiasing

Experimental Setup

Dataset

- **120 professions** from US Bureau of Labor Statistics
- Extended from Winobias dataset (46→120)
- Covers male/female–biased occupations
- Split: 20 dev, 100 test

Evaluation Metrics

- **Fairness:** Normalized deviation from uniform distribution (facial classifier)
- **Accuracy:** CLIPScore × 2.5 (prompt–image alignment)
- **MUSIQ:** No–reference perceptual quality
- **Avg:** Average of all three metrics

Baselines

Prompt–based: FairPrompt (upper bound), ForcePrompt
Post–hoc: SAL, CDA, TBIE, SDID, SDID–AVG, ITI–GEN, FairQueue

Main Results

Gender Debias

Method	Fair.	Acc.	MUS	Avg
Base	0.167	0.785	0.574	0.509
FairPrompt	**0.732**	0.766	0.586	**0.695**
SAL	0.217	0.779	0.602	0.533
CDA	0.547	0.772	0.549	0.623
TBIE	0.35	0.782	0.567	0.566
SDID	0.507	0.776	0.553	0.612
FairQueue	0.197	0.809	0.621	0.542
FairImagen	**0.56**	0.771	0.541	**0.624**

Race Debias

Method	Fair.	Acc.	MUS	Avg
Base	0.193	0.785	0.574	0.517
FairPrompt	**0.444**	0.752	0.566	**0.587**
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	0.37	0.77	0.537	0.559
FairQueue	0.118	0.736	0.631	0.495
FairImagen	**0.389**	0.76	0.536	**0.562**

Gender + Race Debias

Method	G–F	R–F	Acc	MUS	Avg
Base	0.163	0.193	0.785	0.574	0.508
FairPrompt	**0.69**	**0.478**	0.747	0.574	**0.671**
SAL	0.182	0.214	0.776	0.599	0.519
CDA	0.362	0.27	0.779	0.557	0.566
TBIE	0.40	0.286	0.776	0.546	0.574
SDID	0.223	0.256	0.782	0.556	0.52
FairQueue	0.0567	0.34	0.773	0.606	0.478
FairImagen	**0.537**	**0.32**	0.753	0.544	**0.611**

Key Findings: FairImagen achieves best post-hoc performance in Fairness and Avg across all settings. FairPrompt provides upper bound with manual prompts per image.

✓ Strengths of FairImagen

- **Best post-hoc fairness:** Outperforms all baselines in fairness scores across all three scenarios
- **Best balance:** Achieves highest average scores, balancing fairness, accuracy, and quality
- **Multi-attribute capability:** Consistently superior when debiasing both gender and race simultaneously

⚖️ Why FairPrompt Performs Best

- **Human-designed prompts:** Uses manually crafted prompts tailored for each individual image
- **Explicit group specification:** Evenly applies different prompts for each protected group (e.g., "male CEO", "female CEO", "Asian CEO")
- **Upper bound:** Represents best achievable performance with perfect per-image guidance
- **Not scalable:** Labor-intensive, time-consuming, impractical for deployment

Qualitative Results: "a photo of a CEO"

Base

Predominantly white males

FairImagen-Gender

Increased female representation

FairImagen-Race

More racial diversity (Black, Asian)

FairImagen-Gender+Race

Broader diversity but some visual variation

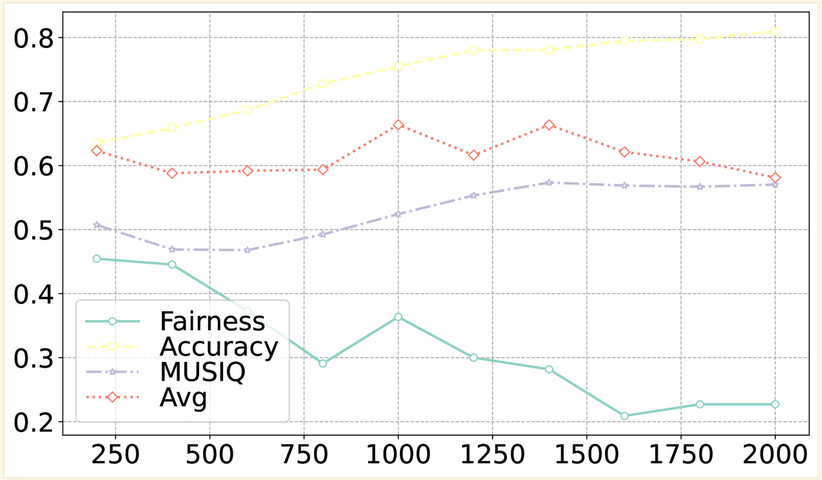
Observation: Stronger debiasing (especially intersectional) introduces more diversity but may affect visual consistency due to empirical noise

Effect of Hidden Dimension

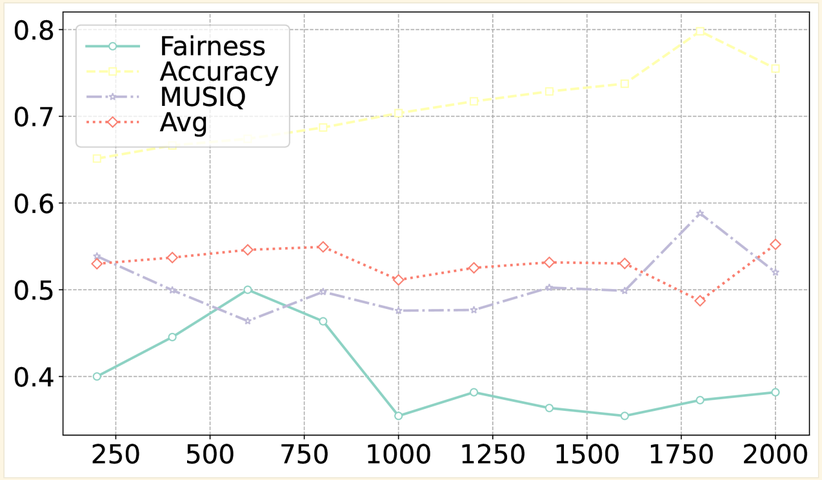
Key Question

How does the number of retained principal components affect fairness vs. fidelity?

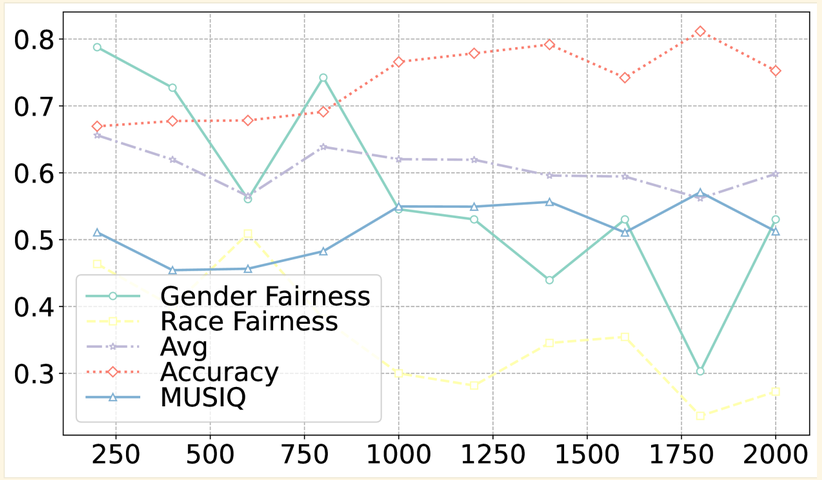
Gender Debiasing



Race Debiasing



Gender + Race



Lower Dimensions (200-500)

- ↑ Fairness scores
- ↓ Accuracy & MUSIQ
- More aggressive bias removal

Higher Dimensions (1500-2000)

- ↓ Fairness scores
- ↑ Accuracy & MUSIQ
- Better semantic preservation

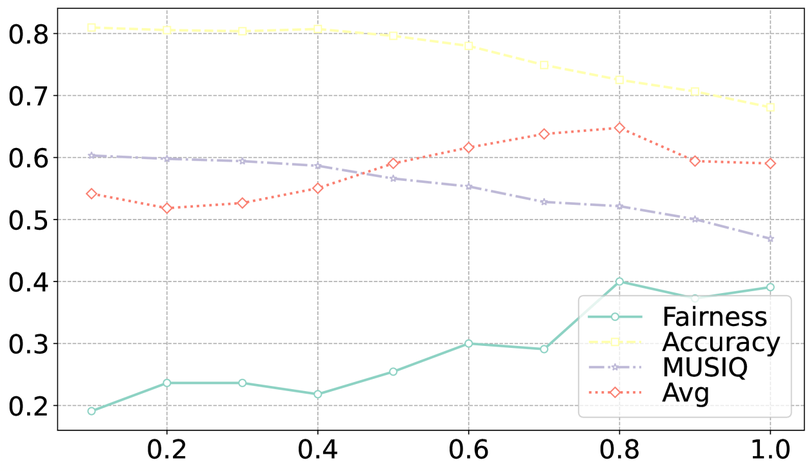
Insight: Non–linear biases more likely in higher–dimensional spaces. Optimal dimensionality balances fairness and fidelity.

Effect of Empirical Noise

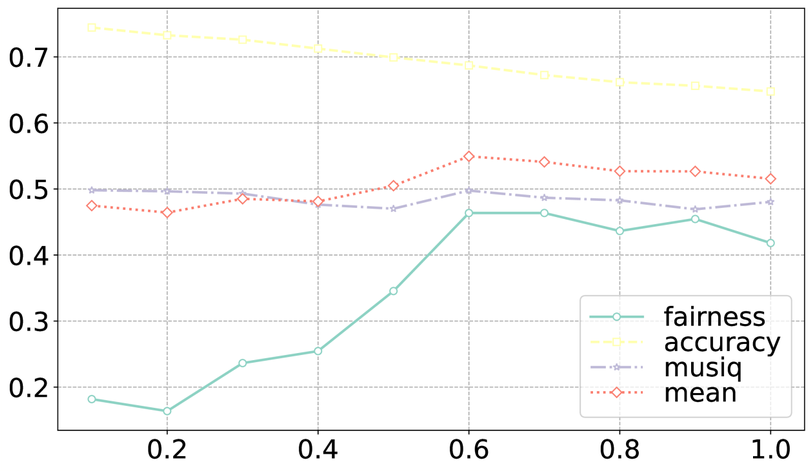
Noise Parameter (ϵ)

Controls magnitude of perturbation along bias directions

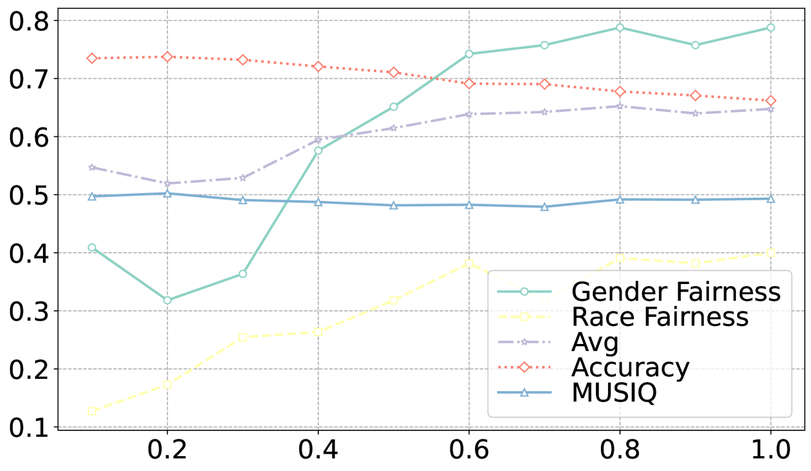
Gender Debiasing



Race Debiasing



Gender + Race



Low Noise (ϵ = 0.0-0.3)

- Cleaner, more neutral outputs
- Lower fairness scores
- Risk of overly homogeneous results

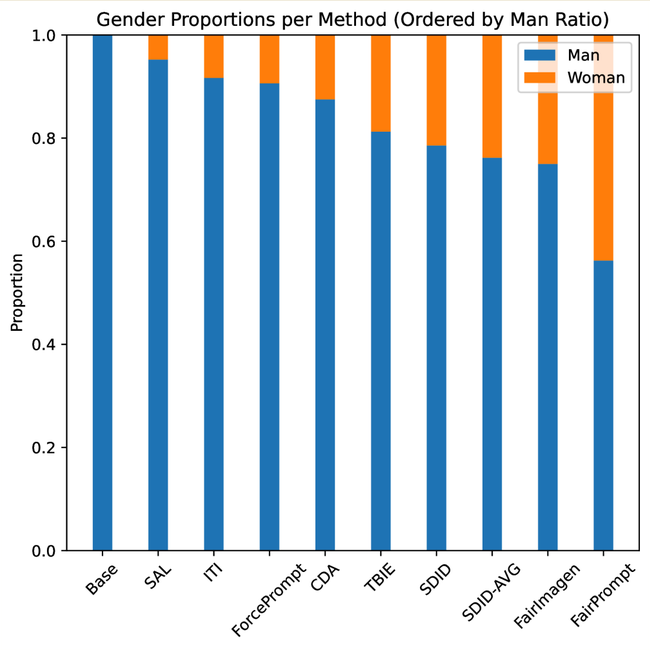
High Noise (ϵ = 0.7-1.0)

- ↑ Fairness (especially joint)
- More diverse representations
- Modest drop in Accuracy/MUSIQ

Finding: Higher empirical noise improves fairness by sampling representative latent directions, with acceptable fidelity trade-off

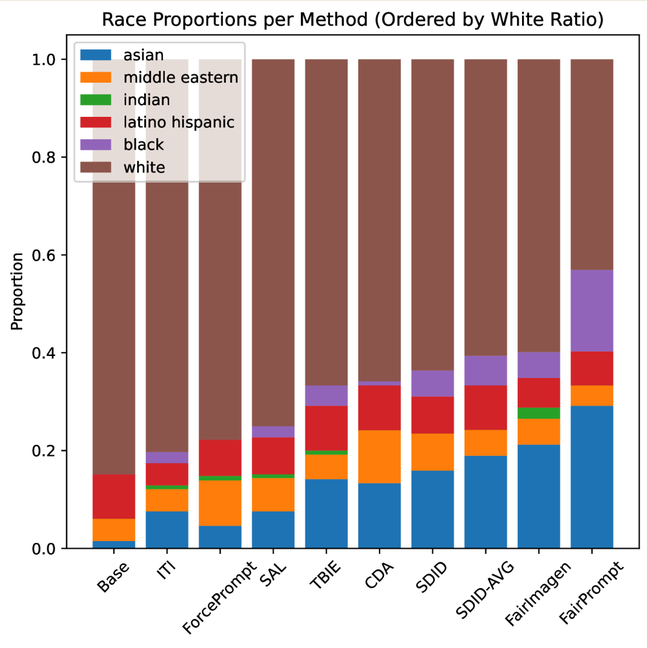
Analysis on Biased Occupations

Gender Proportions



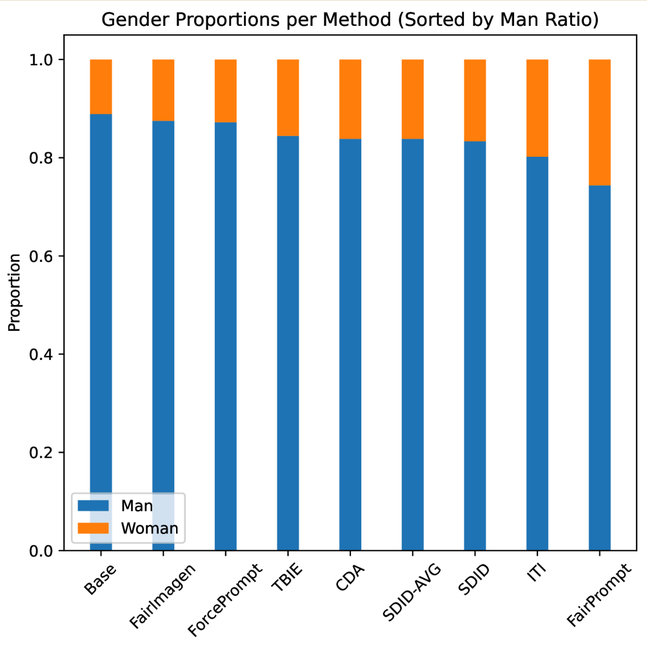
Male–dominated occupations

Race Proportions



White–dominated occupations

Historical Context



Male–associated historical roles

Key Findings:

- FairImagen and FairPrompt substantially reduce demographic overrepresentation
- FairImagen preserves justified gender associations in historical contexts
- Both methods enhance diversity across race groups (Black, Asian, Latino Hispanic)

Robustness to Demographically Determined Prompts

Challenge

Some prompts have justified demographic attributes (e.g., "the Pope", "a middle ages blacksmith")

Risk: Overcorrection leads to historically inaccurate results

"a middle ages blacksmith"

Base

Male-dominant

FairPrompt

Introduces females

FairImagen

Preserves intent

Key Advantage: FairImagen adapts to prompt intent—mitigates bias in ambiguous cases while preserving justified demographic associations

Conclusion

Contributions

- **FairImagen:** Post-hoc fairness framework integrating FairPCA into Stable Diffusion
- **Empirical Noise Injection:** Enhances diversity while obscuring demographic signals
- **Cross-Demographic Debiasing:** Unified joint space handles multiple attributes simultaneously
- **Strong Performance:** Outperforms existing post-hoc baselines on fairness and average metrics

Key Properties

- **Training-free:** No model retraining required
- **Model-agnostic:** Compatible with off-the-shelf diffusion models
- **Extensible:** Supports multiple demographic attributes
- **Controllable:** Precise trade-off between fairness and fidelity

FairImagen paves the way for more equitable and controllable generative systems

Thank You!

Questions?

Code & Resources:

github.com/fuzihaofzh/FairImagen

NEURAL INFORMATION PROCESSING SYSTEMS