Posterior Mean and Initial Noise Guidance for Gender Bias Optimization

Dain Kim, Jinseo Kim, and Sungyong Baik
School of Data Science, Hanyang University

Background

Modern text-to-image models frequently exhibit gender stereotypes when generating images from neutral profession prompts.

Such distortions reduce the reliability and fairness of generated images, highlighting the need for methods that can mitigate bias without retraining the model.

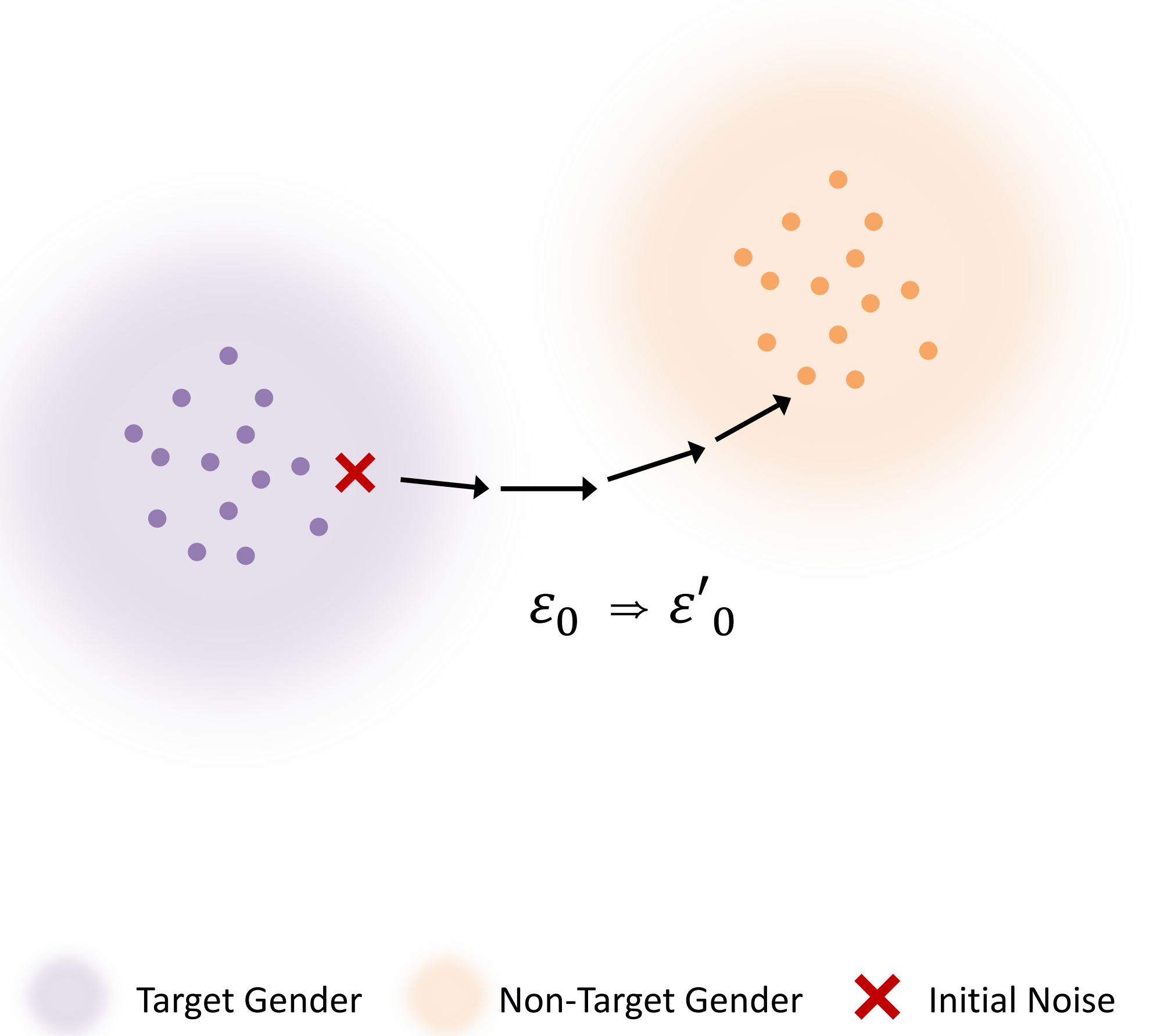
Objective

- Reduce gender bias in **neutral prompts**
- Maintain high-level image quality
- Develop a **training-free** debiasing method

Methods

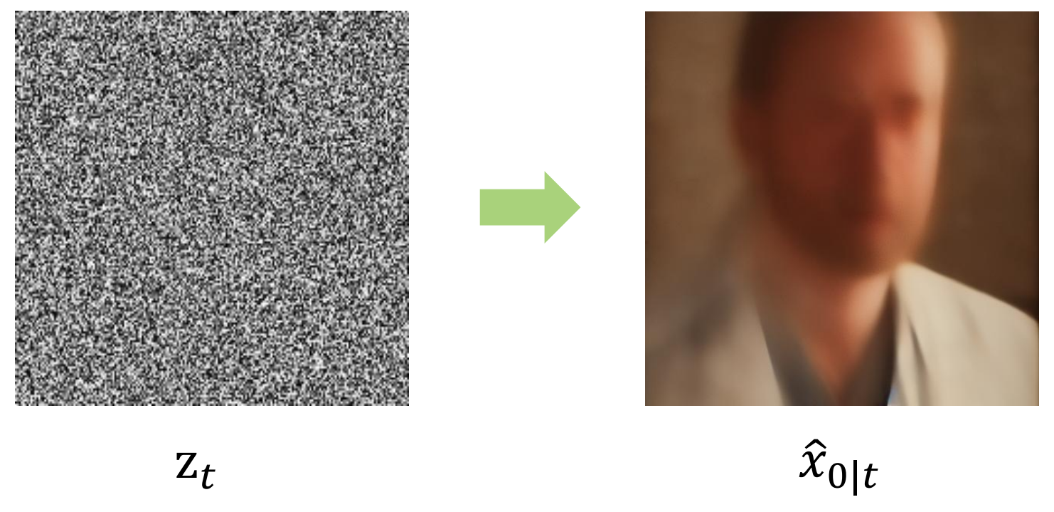
Step 1: Initial Noise Optimization

< Intial latent space >



Step 2: Noise Guidance

1. Compute Posterior Mean



$$\hat{x}_{0|t} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \sqrt{1-\alpha_t}\,\epsilon_\phi(z_t, y, t)\right)$$

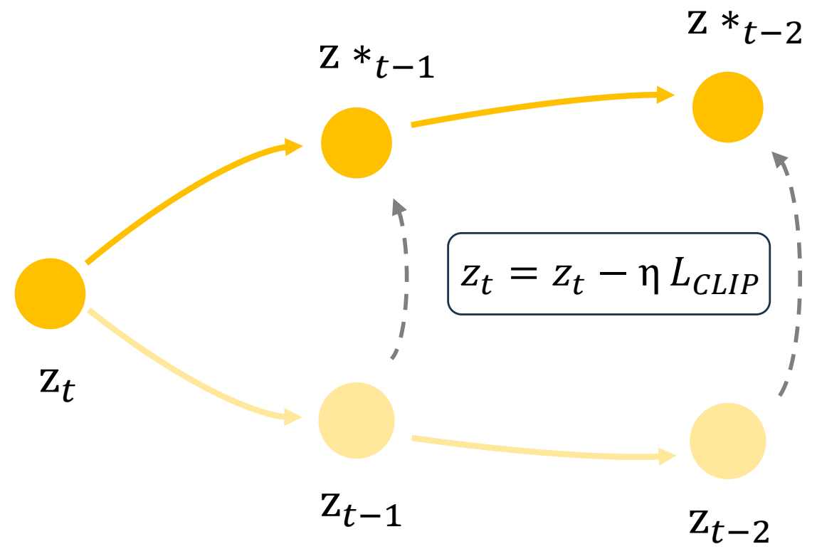
Posterior mean approximates the image implied by the current time step

2. CLIP Score Measurement

A photo of a {gender}

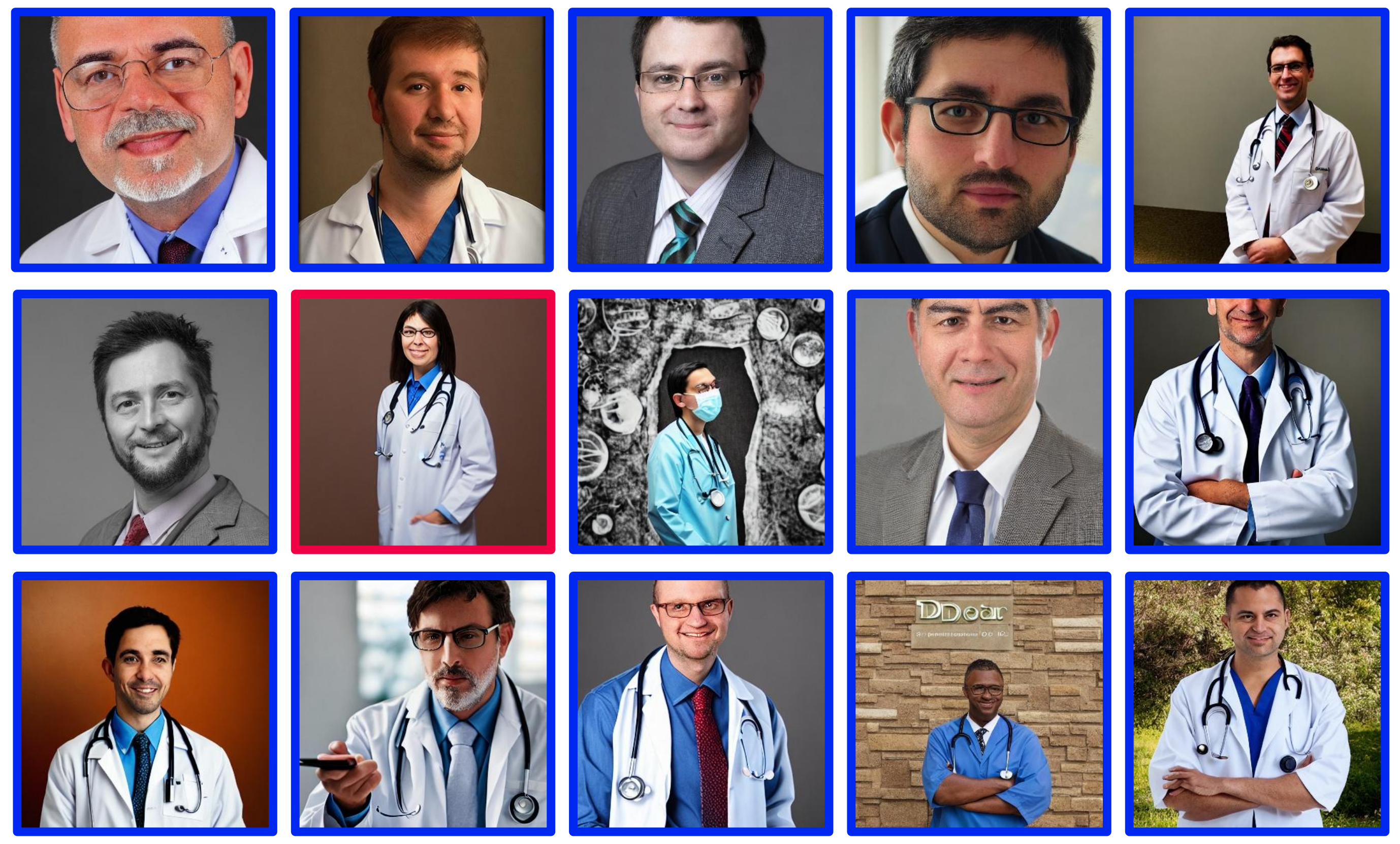
$$\text{CLIPsim}(\hat{x}_{0|t}, y_{target})$$

3. Gradient Step

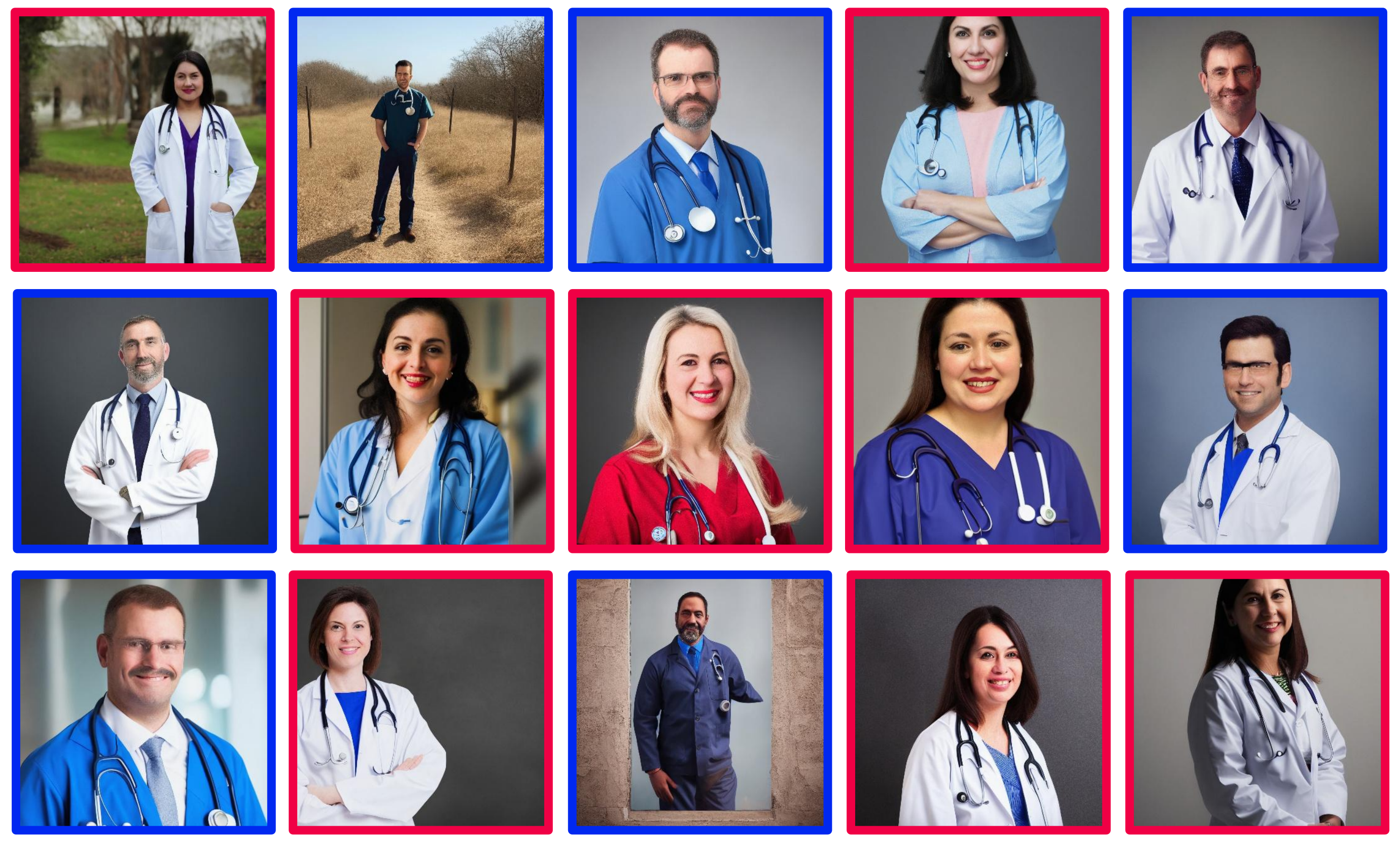


Each step adjusts the latent using CLIP gradients, gradually shifting the trajectory away from biased directions

Results



Stable Diffusion (baseline)



Stable Diffusion (ours)

Conclusions

Our study shows that inference-time optimization can reduce gender bias without modifying the model. This training-free approach fits easily into standard diffusion pipelines and yields more balanced gender representations while preserving image fidelity.

Future work includes extending fairness to attributes such as age, race, and skin tone. We also plan to evaluate generalization across different backbones and prompt domains.