CVPR Nashville JUNE 11-15, 2025

Dissecting and Mitigating Diffusion Bias via Mechanistic Interpretability

Yingdong Shi[1*], Changming Li[1*], Yifan Wang[2], Yongxiang Zhao[1], Anqi Pang[3], Sibei Yang[1], Jingyi Yu[1], Kan Ren[1†]

[1]ShanghaiTech University [2]Stony Brook University [3]Tencent PCG

{shiyd2023, lichm2024, renkan}@shanghaitech.edu.cn

[*]Equal contribution [†]Corresponding author

上海科技大学 ShanghaiTech University Stony Brook University Tencent 腾讯

Diffusion models often perpetuate social biases, including gender, age, and race.

These biases may lead to detrimental effects in real-world contexts, such as reinforcing stereotypes in media representations or perpetuating inequalities in automated decision-making systems.

Existing approaches to debias diffusion models generally fall into two main strategies.

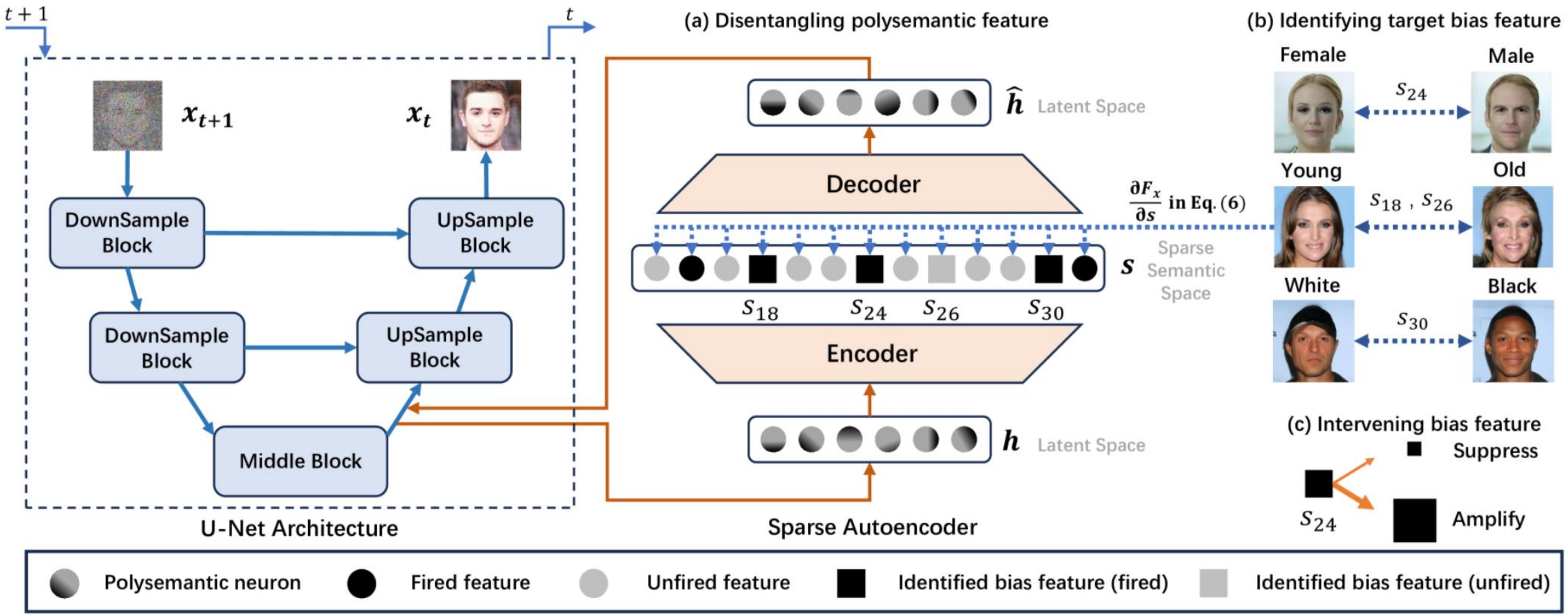
- *Train* unbiased models from scratch on biased datasets [1] or *fine-tune* the target model [2]
    - resource-intensive
    - significantly impact the model's original performance
- *Guide* or *edit* the generation process [3, 4, 11]
    - use gradients w.r.t distribution loss or learn a latent vector to guide the generation process
    - risking overcorrecting the model behavior or distorting non-target attributes

We are the first to interpret and find the *mechanisms* (features) related with property of generated contents, specifically those cause the bias output, within the semantic space in diffusion models.

Question: can we find mechanisms that help to better understand and mitigate bias within diffusion models?

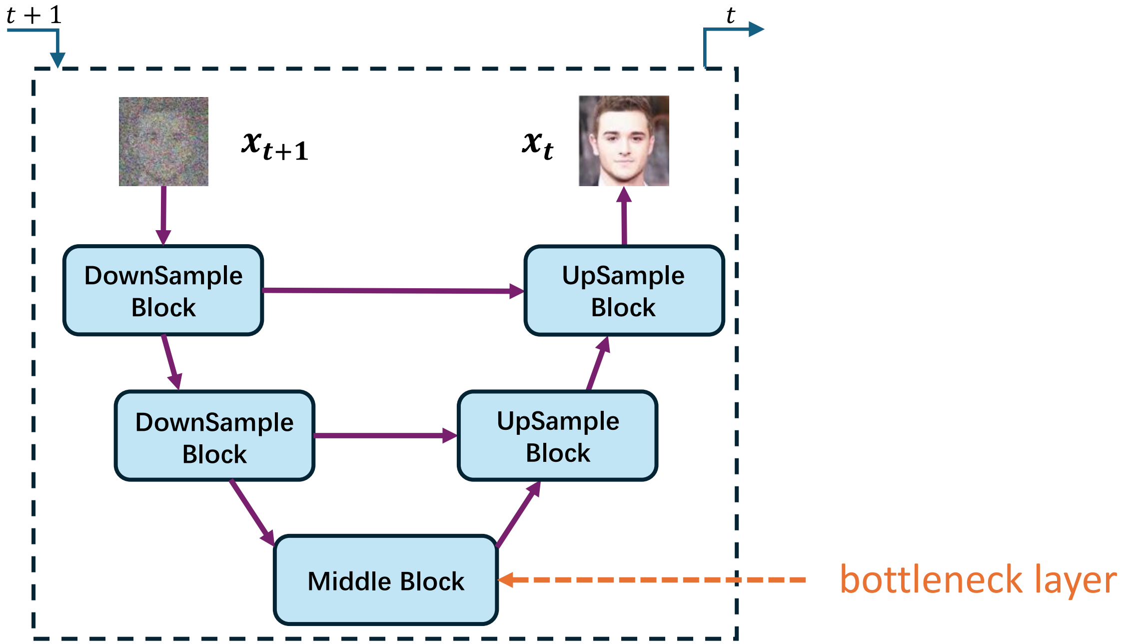
Methodology

(a) Disentangling polysemantic feature (b) Identifying target bias feature (c) Intervening bias feature



Methodology

We follow [4] target at bottleneck layer, which is a semantic space in the U-Net [7] of diffusion models, suggesting that bias related mechanism may concentrate in this space.



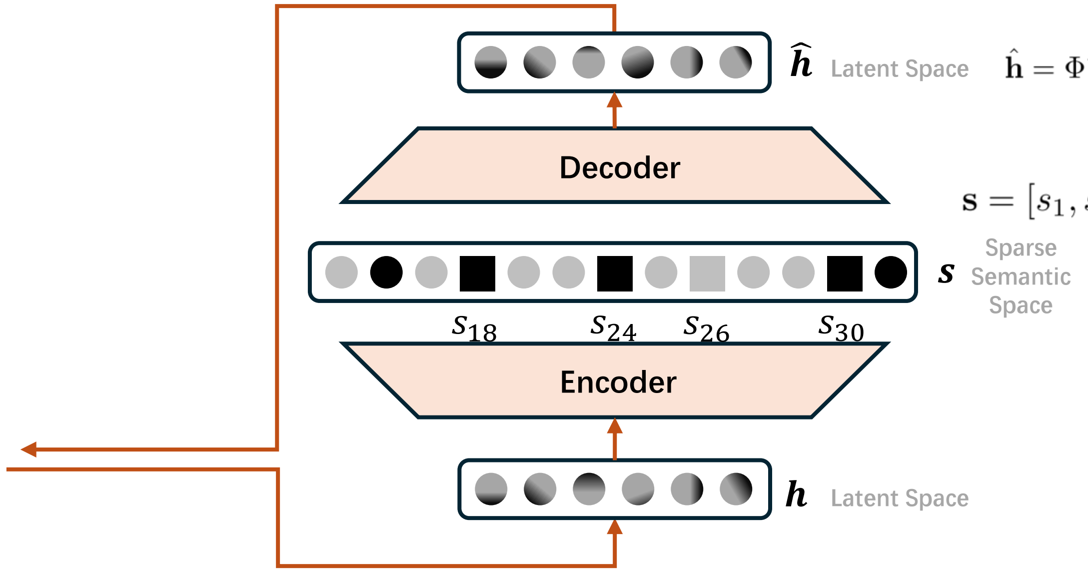
U-Net Architecture

Methodology

Latent space is known as *ploy-semantic* [6], which means one neuron links to multiple unrelated concepts.

We follow [7] that finds interpretable features using Sparse Autoencoders to disentangle the latent space from the model layer into a *mono-semantic* space.

(a) Disentangling polysemantic feature



$$\hat{\mathbf{h}} = \Phi^{-1}(\mathbf{s}) = \mathbf{W}_{\text{dec}}\mathbf{s} + \mathbf{b}_{\text{pre}}$$

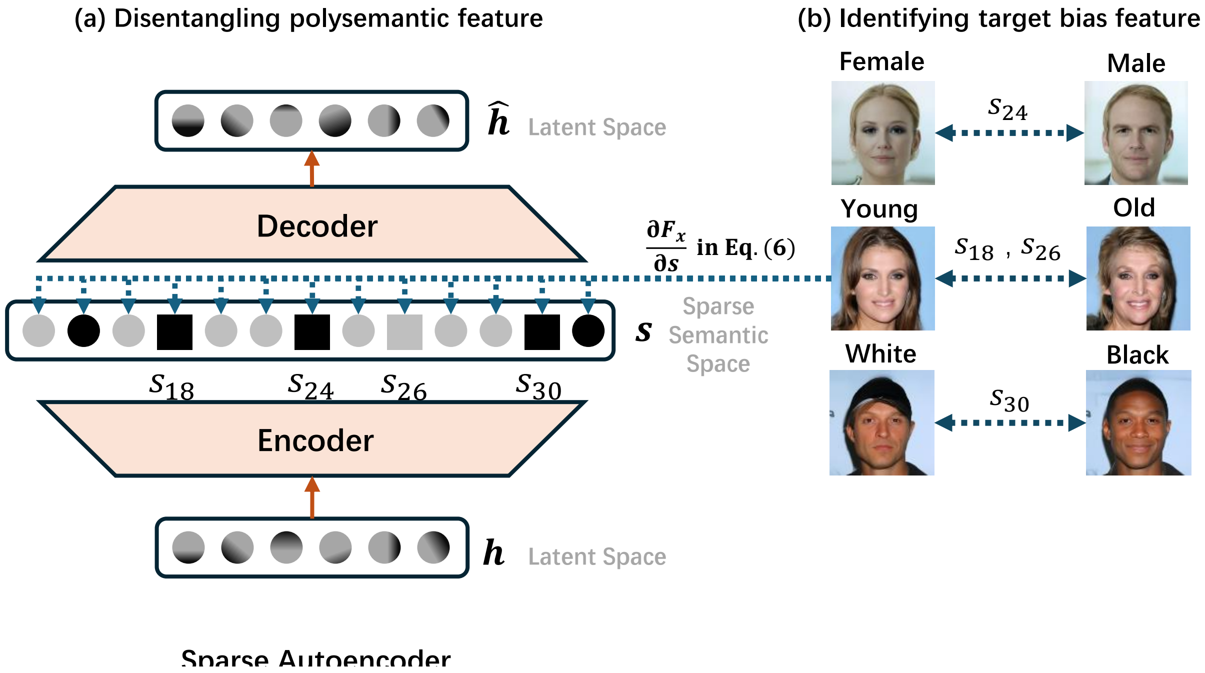
$$\mathbf{s} = [s_1, s_2, \ldots, s_m] = \Phi(\mathbf{h}) = \mathrm{TopK}(\mathbf{W}_{\text{enc}}(\mathbf{h} - \mathbf{b}_{\text{pre}}))$$

Sparse Autoencoder

Methodology

Social biases are related with activations of bias features.

To identify which feature causes bias generations, we train a light-weight classifier F_x , and then we identify bias related features using a gradient-based attribution method.

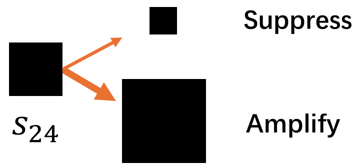


$$S(s_i;\mathbf{x}) = (s_i - s_i') \cdot \int_{\alpha=0}^{1} \frac{\partial F_{\mathbf{x}}(\mathbf{s}' + \alpha(\mathbf{s} - \mathbf{s}'))}{\partial s_i} \, d\alpha$$

Methodology

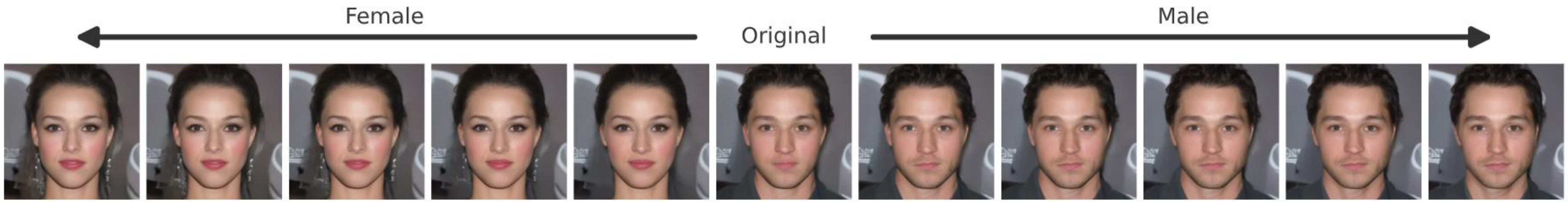
The bias level in generated images exhibits a monotonic relationship with the scale of bias-related features. Thus, we control over the level of bias in generated content by adjusting the activation scale of these features.

(c) Intervening bias feature



$$s_i = \text{Intervene}(s_i) = \begin{cases} \beta s_i & (\text{Scaling}), \text{ or} \\ s_i + \beta & (\text{Adding}). \; \forall i \in \mathbf{A}, \beta \in \mathbb{R} \end{cases}$$

Intervention results



Results

Experiment settings

- Diffusion Model Architecture
  - P2 model [8]
  - Stable Diffusion v1.5 [9]
- Baselines
  - interpretability methods
    - Activation [10] locates influential neurons according to the neuron activation values.
    - Latent Direction [11] identifies interpretable semantic directions within the latent space in the U-Net.
  - guidance-based methods
    - Latent Editing [4] tries to learn a latent vector to guide the unbiased generation within the bottleneck layer of U-Net.
    - H-Distribution [3] employs distributional loss on bottleneck layer as guidance in diffusion models.
  - Finetuned-based methods
    - Finetuning [2] finetunes the model to align with a user-defined target distribution.
- Evaluation metrics
  - Fairness Discrepancy (FD) [3] which calculates the Euclidean distance between a reference distribution and the bias distribution of generation.
  - Fréchet Inception Distance (FID) [12] measures quality of generated images.
  - CLIP-T calculates the CLIP-based semantic similarity between the generated image and the input text prompt.
  - CLIP-I assesses the similarity between originally generated images and images after debiasing.

Results

Question: How effectively does DiffLens mitigate social bias while maintaining image quality?

Finding 1: DiffLens effectively neutralizes biases while preserving image quality.

P2 model [8]

Category	Method	Gender (2)			Age (3)			Race (4)		
		FD ↓	FID ↓	CLIP-I ↑	FD ↓	FID ↓	CLIP-I ↑	FD ↓	FID ↓	CLIP-I ↑
	Original	0.226	33.38	-	0.592	33.38	-	0.718	33.38	-
Guidance-based	Latent Editing [32]	0.003	**29.85**	0.9474	0.606	33.71	0.9092	**0.317**	34.61	0.9081
	H-Distribution [43]	0.048	31.31	0.9440	0.511	34.21	0.8594	0.494	36.19	0.8762
Interpretability	Activation [6]	0.190	34.27	0.8060	0.544	48.91	0.7793	0.700	46.13	0.7846
	DIFFLENS (Ours)	**0.002**	31.93	**0.9479**	**0.401**	**31.71**	**0.9414**	0.447	**33.47**	**0.9111**

- Stable Diffusion v1.5 [9]

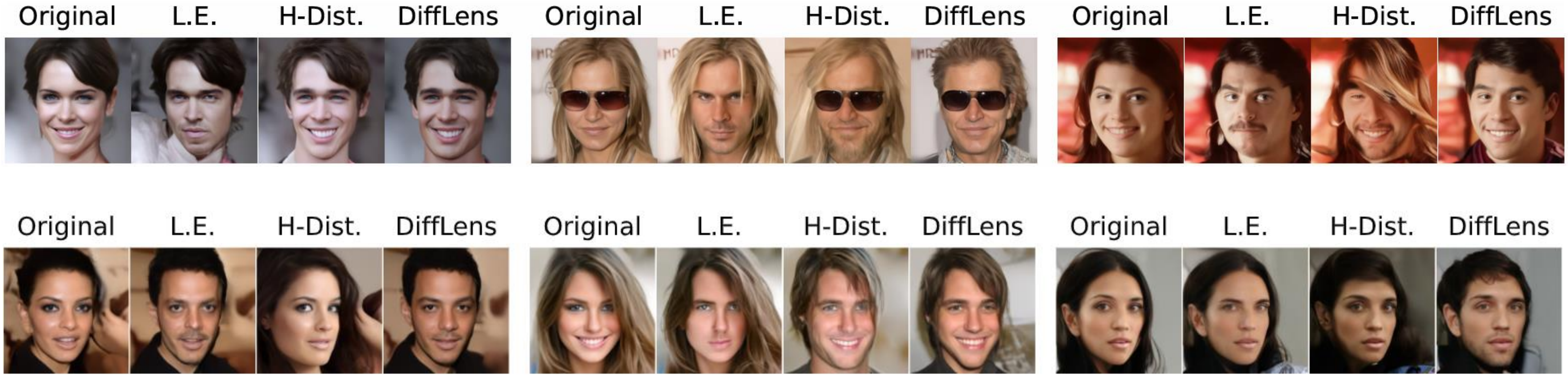
Method	Gender (2)				Age (3)				Race (4)			
	FD ↓	FID ↓	CLIP-I ↑	CLIP-T ↑	FD ↓	FID ↓	CLIP-I ↑	CLIP-T ↑	FD ↓	FID ↓	CLIP-I ↑	CLIP-T ↑
Original	0.564	120.06	-	0.6155	0.752	120.06	-	0.6155	0.558	120.06	-	0.6155
Latent Editing [32]	0.408	166.11	0.8253	0.6005	0.682	200.90	0.8527	**0.6122**	0.524	153.05	0.8804	0.6086
H-Distribution [43]	0.222	151.68	0.8475	0.6087	0.506	147.71	0.8345	0.6098	0.544	126.90	0.8255	0.6100
Latent Direction [33]	0.305	129.37	0.8058	0.6091	0.052	113.81	0.8151	0.6067	**0.175**	128.30	0.8211	0.6132
Fintuning [54]	0.050	161.47	**0.8779**	**0.6095**	0.746	161.47	**0.8779**	0.6095	0.198	161.47	0.8779	0.6095
DIFFLENS (Ours)	**0.046**	**112.83**	0.8501	0.6090	**0.049**	**99.17**	0.8778	0.6057	0.401	**119.86**	**0.9096**	**0.6149**

FD stands for Fairness Discrepancy, **FID** measures image quality, **CLIP-I** assesses the similarity between originally generated images and images after debiasing, **CLIP-T** calculates the CLIP based semantic similarity between the generated image and the input text prompt

Results

Question: Can DiffLens accurately identify intrinsic mechanism of bias generation?

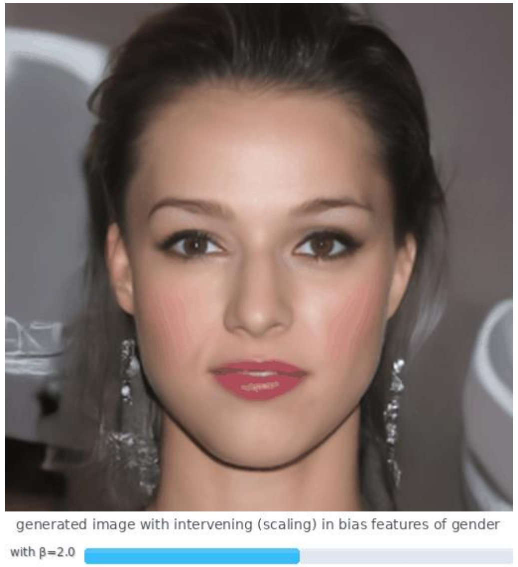
Finding 2: DiffLens preserves overall image semantics.



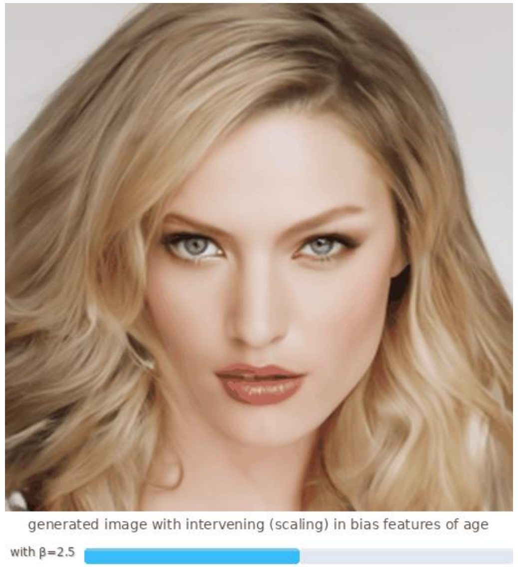
Results

Question: How well does DiffLens control the bias level through intervening in the found bias features?

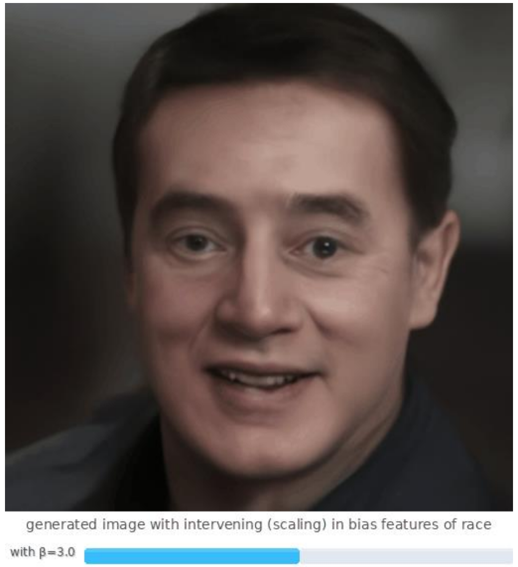
Finding 3: DiffLens produces natural transitions, consistently preserving semantic feature.



Female-Male



Young-Old



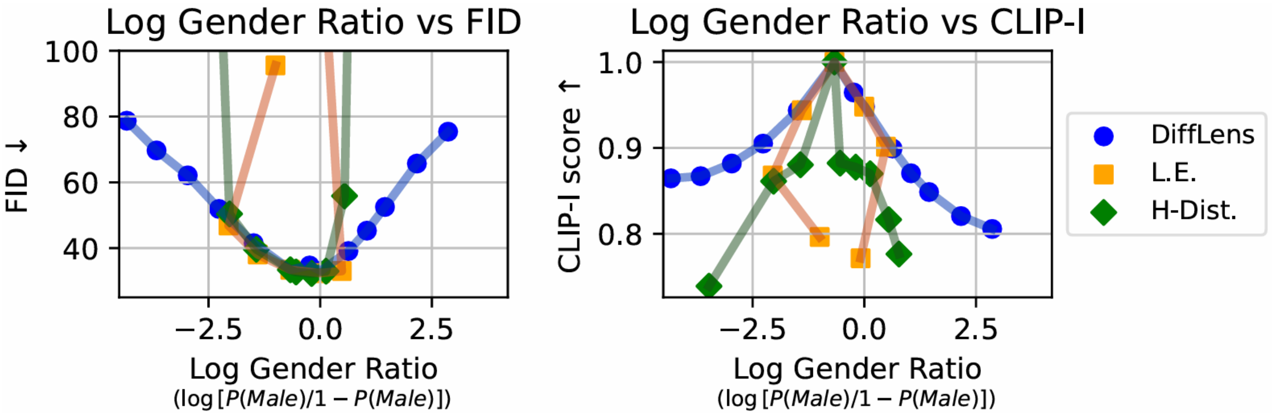
Asian-White-Black

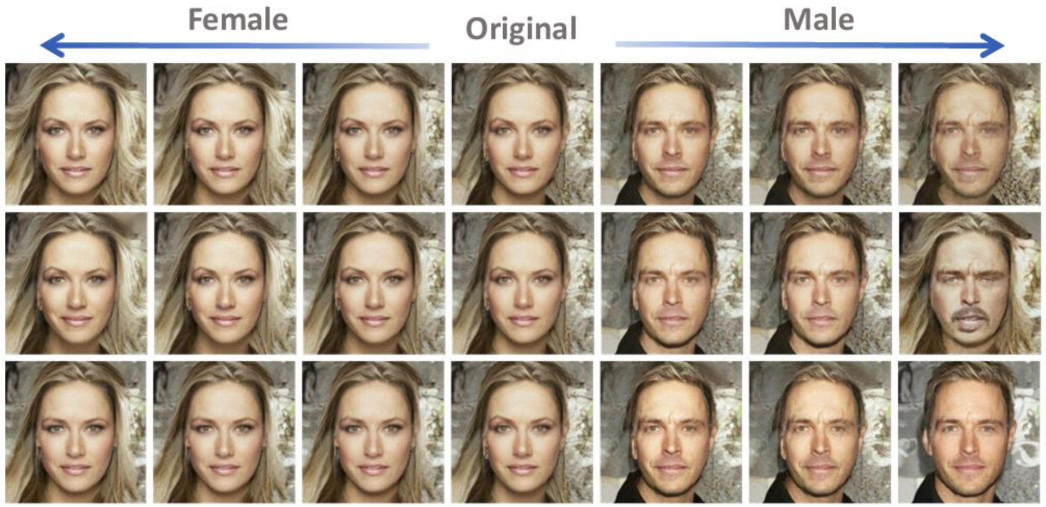
Results

Question: How well does DiffLens control the bias level through intervening in the found bias features?

Finding 4: DiffLens offers broader bias control.

- The Log Gender Ratio in the right reflects the log Of male to female ratio in the generated images, with 0 indicating balance.
- We achieve overall stable results.

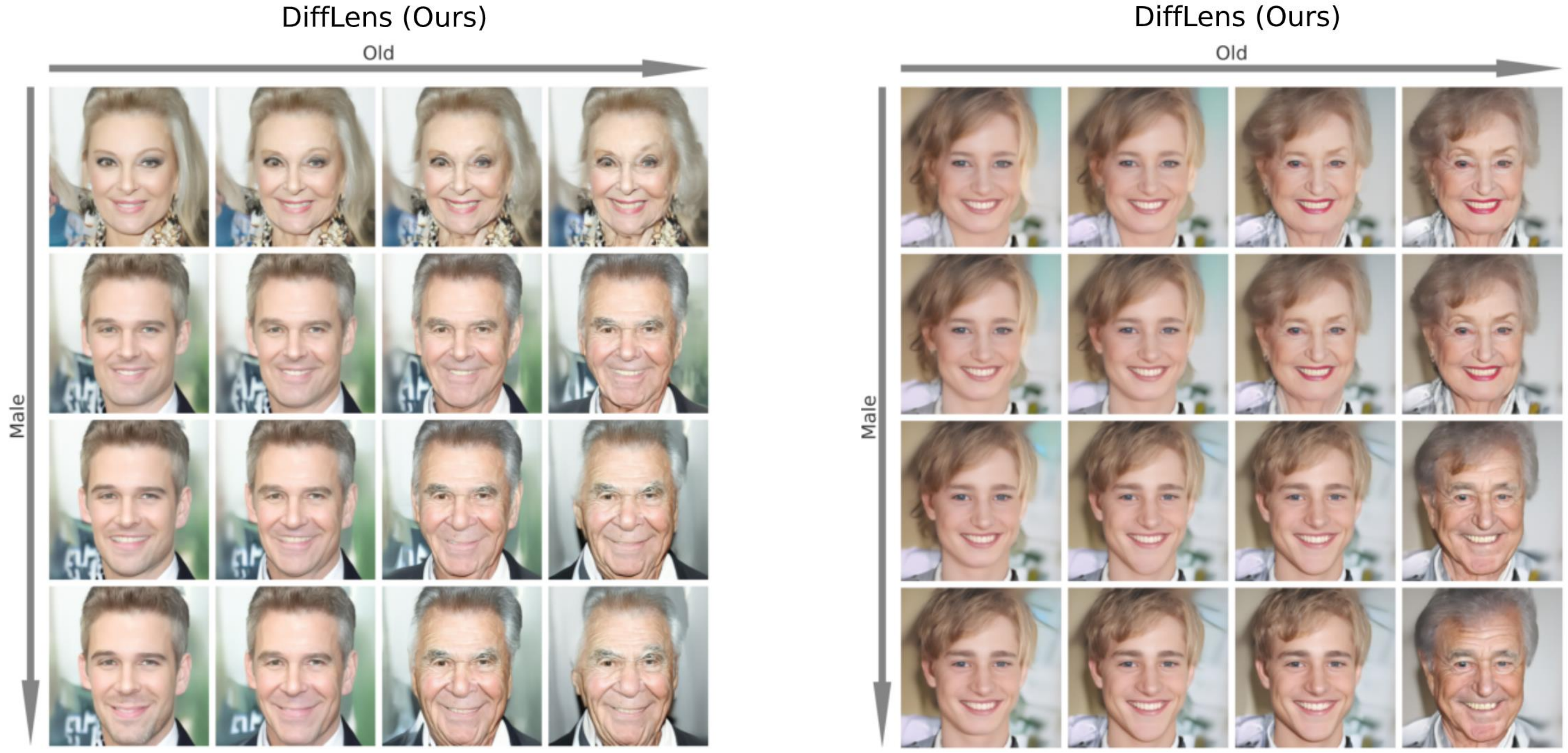




Results

Question: Can DiffLens control multi-attributes simultaneously?

Finding 5: DiffLens is able to disentangle different bias features and accurate identification of these features.



Reference

[1] Yeongmin Kim, Byeonghu Na, Minsang Park, JoonHo Jang, Dongjun Kim, Wanmo Kang, and Il chul Moon. Training unbiased diffusion models from biased dataset. (ICLR 2024).

[2] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. (ICLR 2024).

[3] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. (CVPR 2024).

[4] Mingi Kwon, Jaeseok Jeong, and Young jung Uh. Diffusion models already have a semantic latent space. (ICLR 2023)

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. (MICCAI 2015)

[6] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. (arXiv 2022).

[7] Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. (ICLR 2024).

[8] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. (CVPR 2022)

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz,Patrick Esser, and Bj¨orn Ommer. High-resolution image synthesis with latent diffusion models. (CVPR 2022)

[10] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron- explainer/paper/index.html, 2023

[11] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jin-dong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. (CVPR 2024)

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. (NeurIPS 2017).

Key takeaways

- We introduced a novel approach DiffLens, for mitigating social biases in diffusion models by dissecting, analyzing and intervening in the internal mechanisms of the diffusion model.
- We disentangle, identify, and control bias-related features with precision, allowing targeted bias mitigation while preserving non-target attributes.
- We offer an interpretable solution to bias mitigation in diffusion models.

Our paper and code are available at project page