GENIE: Socially Unbiased Generative Text-to-Image Editing

Julia K. Lau†
Monash University, Malaysia campus
julia.lau@monash.edu

Raphaël C.-W. Phan
Monash University, Malaysia campus
raphael.phan@monash.edu

Sailaja Rajanala
Monash University, Malaysia campus
sailaja.rajanala@monash.edu

Ingemar J. Cox⋆
University College London, UK
ingemar@ieee.org

Arghya Pal
Monash University, Malaysia campus
arghya.pal@monash.edu

***Abstract*—Generative diffusion models often exhibit societal biases in sensitive personal attributes such as age, gender, and race. In this work, we describe GENIE – a method to reduce such biases in a variety of classifier-free diffusion models used for image editing. Our method implicitly incorporates debiasing terms together with the user's explicit edit instruction to reduce bias. This automatic method relieves the user from needing to modify edit instructions in order to avoid bias. Further, no additional training is needed. Experimental results are provided based on modifications to four diffusion models, namely InstructPix2Pix, Stable Diffusion 1.5, Stable Diffusion 2.1, and Stable Diffusion XL. We show that, on average, bias is reduced by 31% in gender, 15% in age, 39% in race.**

***Index Terms*—Societal Biases, Text-to-Image, Debiasing.**

I. Introduction

Diffusion models have emerged as a popular class of generative models. These models have generative capabilities in a diverse range of tasks, from image editing to image creation. Correspondingly, diffusion models are increasingly being adopted by the public. For instance, OpenAI [1] has integrated DALL-E 3 [2] into ChatGPT [3], enabling ChatGPT to create images based on users' textual descriptions [4]. Despite the learning and generative capabilities of these models, researchers have shown that these models suffer from human-like biases and stereotypes [5]–[7]. As these models are being employed in large-scale applications, such biased outputs could reinforce existing societal stereotypes. The biggest contributor to these biases is the datasets that these models are trained on, which are commonly obtained from the internet [8], [9]. The models then learn the implicit biases present in the data and reflect these biases in the generated images.

To reduce biases, we propose a guidance-based method to debias text-to-image editing models during inference, which requires no training. Our approach leverages classifier-free diffusion guidance (discussed in Sec. III), to encourage the diffusion model to generate unbiased content. The key advantage of our model is that it does not require any expensive model training nor fine-tuning. Furthermore, we demonstrate that our debiasing method can be extended to pre-trained generative models that utilise classifier-free diffusion guidance, such as Stable Diffusion (SD) and InstructPix2Pix (IP2P).

We summarise our main contributions as follows:

- We introduce GENIE, a guidance-based approach to mitigate biases in image editing models during inference.
- We show that GENIE requires no fine-tuning and can be integrated with classifier-free guided generative models.
- Our results show that GENIE can achieve an average bias reduction of up to 31% for gender bias, 15% for age bias, and 39% for race bias, averaged over the four models we considered.

II. Related work

A. Social Biases in Text-to-Image Models

Numerous studies have investigated biases in text-to-image generative models, revealing that these models encode a broad range of biases [5], [6], [10]–[14]. Common findings highlight race and gender imbalances, with textual prompts of occupations like *software developer* often generating images of masculine, fair-skinned individuals and prompts like *attractive* generating Caucasian-looking faces as opposed to Asian or African faces [6], [12], [15]. Recent work [16] proposed BiasPainter to extend bias analysis to text-to-image editing models. BiasPainter used diverse seed images of people and applied neutral text prompts to the models. The edited images were compared with the original seed images to examine changes related to the sensitive attributes of gender, race, and age. Ideally, these attributes should remain unchanged when neutral prompts are used, but the study revealed that such prompts often triggered unexpected changes to these attributes, indicating biases in the models.

B. Debiasing Text-to-Image Models

Some methods have been proposed to address biases. A study [17] reduced biases in image editing models that leveraged Contrastive Language-Image Pretraining (CLIP). They showed that CLIP embeddings learned associations between professions and gender or race. To mitigate this, they proposed a text-based bias mitigation technique to remove gender subspaces from CLIP text embeddings. However, the authors found that text-based debiasing failed to preserve the input

† Julia's research is funded in part by the HUMANΘID project.
⋆ Part of this work done while at Monash as an Elite Visiting Professor.

image's identity hence they also proposed a gradient-based latent code optimisation to minimise identity loss. Unfortunately, what can be observed in their debiased images is that the original faces are essentially transferred onto the output images, ensuring that both the input and output share the same attributes. While forcing the same face on the output reduces the biases, this severely constrains the flexibility and the range of outputs the generative models can produce.

Fair Diffusion [18] extends classifier-free guidance by adding a fair guidance term to address biases in image generation models. The generation process is guided towards both the text prompt and fairness instruction simultaneously, with the fairness instruction defined using textual descriptions of specific attributes. To promote fairness, the authors randomly sampled attributes from a desired probability distribution, ensuring a diverse representation in the generated images. Safe Latent Diffusion [19] also used a guidance-based approach to steer models, aiming to prevent inappropriate content in text-to-image generation. They introduced a safety guidance in classifier-free diffusion guidance, where textual inappropriate concepts are defined and text conditioning is used to guide the model away from inappropriate concepts. However, both approaches are limited to image generation models, and have not been extended to editing models, which require an input image along with a text prompt. The key difference lies in the additional image input and the need to preserve its sensitive attributes to reduce model bias. While debiasing techniques for image generation models can be helpful, they fail to consider the sensitive attributes of the input image. Hence, it is essential to extend existing debiasing techniques to image editing.

Different from existing work, GENIE aims to debias text-to-image **editing models** using a **guidance-based approach**. To the best of our knowledge, GENIE is the first to apply this approach to debias image editing models.

III. Methodology

A. Classifier-Free Diffusion Guidance

Diffusion models [20] are trained to generate data samples through a series of denoising autoencoders that estimate the score of a data distribution, which points towards higher data density. Classifier-free diffusion guidance [21] is a conditioning mechanism that trades diversity for fidelity in images generated by a diffusion model. Diversity refers to its ability to produce a variety of outputs while fidelity indicates how closely the generated image aligns with the input conditioning. Unlike classifier guidance [22], classifier-free guidance eliminates the need for a separate classifier by jointly training both unconditional and conditional denoising processes together. During inference, the score estimates are modified such that:

$$\tilde{\epsilon_\theta}(z_t, c_T) = \epsilon_\theta(z_t, \varnothing) + s_T(\epsilon_\theta(z_t, c_T) - \epsilon_\theta(z_t, \varnothing)) \quad (1)$$

where s_T denotes the guidance scale, c_T the text conditioning, z_t the noisy latent, and ϵ_θ the noise estimate. This shifts the unconditioned noise prediction $\epsilon_\theta(z_t, \varnothing)$ towards the text-conditioned $\epsilon_\theta(z_t, c_T)$, with s_T controlling the text's influence.

Classifier-free diffusion guidance can also be extended to take in an additional image conditioning c_I as in IP2P [23], which builds on SD to enable image editing via edit instructions. To edit an image, the unconditioned noise estimate $\epsilon_\theta(z_t, \varnothing, \varnothing)$ can be guided by both an image-conditioned estimate $\epsilon_\theta(z_t, c_I, \varnothing)$, and a combined image and text-conditioned estimate $\epsilon_\theta(z_t, c_I, c_T)$. This allows the final image to be influenced by both the input image c_I and prompt c_T, with s_I and s_T controlling the strength of each input.

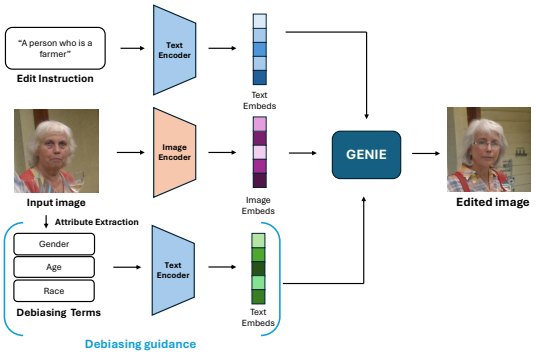

Fig. 1: GENIE. A user inserts an input image along with a neutral edit instruction, both of which are encoded into embeddings. To provide debiasing guidance, we extract textual debiasing terms from the attributes in the input image and encode them into text embeddings. This enables the model to recognise and preserve these attributes in the edited image.

B. Proposed Debiasing Method: GENIE

Our proposed socially-unbiased **GEN**erative text-to-**I**mage **E**diting method (GENIE) introduces a novel debiasing approach for providing *debiasing guidance* to latent diffusion models for image editing, in order to reduce biases present in the generated outputs of editing models. Our approach leverages the flexibility of classifier-free diffusion guidance and the model's knowledge to steer the diffusion process towards a less biased generative space through textual *debiasing terms*. The key aspect in GENIE is the introduction of a textual debiasing term D in addition to a text prompt T. To meet our fairness criteria—where a neutral prompt results in the edited image retaining the attributes of the input image [16] —we must first extract the gender/age/race attributes from the input image. The identified attributes are then used to define the debiasing terms. As a result, we have a total of 3 inputs: text prompt, input image, and debiasing terms, as shown in Figure 1.

We apply GENIE to SD as follows. In the classifier-free diffusion guidance equation in Eq. (1) employed by text-to-image models such as SD, there are two ϵ-predictions: $\epsilon_\theta(z_t, c_T)$ and $\epsilon_\theta(z_t, \varnothing)$. Given the introduction of an additional input in our approach, it is necessary to extend classifier-free diffusion guidance to account for the debiasing term. To achieve this, we adapt Safe Latent Diffusion (SLD) [19]. Our key difference though is that SLD aims to guide image generation models *away* from inappropriateness, whereas our approach focuses on guiding image-editing models *towards* a less biased output. To extend this approach, we now employ

three ϵ-predictions to shift the unconditioned score estimate $\epsilon_\theta(z_t, \varnothing)$ *towards* both the prompt-conditioned $\epsilon_\theta(z_t, c_T)$ and the debiasing term-conditioned estimate $\epsilon_\theta(z_t, c_D)$. This is mathematically formulated as:

$$\begin{aligned}\tilde{\epsilon_\theta}(z_t, c_T, c_D) \quad &= \quad \epsilon_\theta(z_t, \varnothing) + \\ & \quad s_g(\epsilon_\theta(z_t, c_T) - \epsilon_\theta(z_t, \varnothing) + \gamma(z_t, c_D))\end{aligned} \tag{2}$$

where the debiasing guidance term γ is defined as:

$$\gamma(z_t, c_D) = \mu(c_T, c_D; s_D, \lambda)(\epsilon_\theta(z_t, c_D) - \epsilon_\theta(z_t, \varnothing)) \tag{3}$$

with c_D denoting the debiasing term. The function μ considers the elements of the text-conditioned score estimate that would lead the generation process towards the debiasing terms. If the element-wise difference between the text-conditioned score and debiasing-conditioned score is below a set threshold λ, μ scales the difference by a guidance scale s_D. Otherwise, it is set to 0. In other words, we aim to scale up the text-conditioned scores that lie close to the debiasing term, so that the generated image better reflects the desired attributes, as defined by the debiasing term. Hence, $\mu(c_T, c_D; s_D, \lambda)$ =

$$\begin{cases} max(1, |\phi|), & \text{where } \epsilon_\theta(z_t, c_T) \ominus \epsilon_\theta(z_t, c_D) < \lambda \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\phi = s_D(\epsilon_\theta(z_t, c_T) - \epsilon_\theta(z_t, c_D))$ and $\ominus$ denotes element-wise subtraction.

GENIE can also be applied to IP2P type of editing models. Our extension introduces a novel adaptation where the debiasing guidance process must now account for two noise estimates by both text conditioning c_T and image-based conditioning c_I. In this case, the debiasing terms D can be introduced to shift the unconditioned score estimate closer to both the prompt-conditioned and image-conditioned scores, while aligning it with the debiasing term-conditioned score. The extended equation for IP2P-like models is formulated as:

$$\begin{aligned}\tilde{\epsilon_\theta}(z_t, c_I, c_T, c_D) = & \epsilon_\theta(z_t, \varnothing, \varnothing) \\ & + s_I(\epsilon_\theta(z_t, c_I, \varnothing) - \epsilon_\theta(z_t, \varnothing, \varnothing)) \\ & + s_T(\epsilon_\theta(z_t, c_I, c_T) - \epsilon_\theta(z_t, c_I, \varnothing)) \\ & + s_D(\epsilon_\theta(z_t, c_I, c_D) - \epsilon_\theta(z_t, c_I, \varnothing))\end{aligned} \tag{5}$$

where $\tilde{\epsilon_\theta}(z_t, c_I, c_T, c_D)$ is the new modified noise prediction. There is an additional debiasing term-conditioned noise estimate $\epsilon_\theta(z_t, c_I, c_D)$ to guide the model's output towards the debiasing attributes; s_D controls the debiasing strength.

One may ask why a user does not explicitly restrict the generative model to enforce a specific gender/race/age by specifying a gender/race/age word in the text prompt, i.e. use a non-neutral prompt? While this may reduce bias, it requires the user to be aware of potential biases and to make a conscious effort to address them. Moreover, this does not resolve the fact that a neutral prompt would still lead to biased outputs. Furthermore, by maintaining the debiasing term as a distinct guidance component, we make a clear distinction between the edit instruction, and the debiasing term. This separation allows greater flexibility in adjusting the strength of the debiasing effect without affecting the semantics of the editing prompt – we can control the scale of the debiasing guidance term to fine-tune how much bias mitigation is applied. For instance, setting the scales (μ in Eq. 3 and s_D in Eq. 5) to 0 disables the debiasing guidance, while increasing them enhances the debiasing strength.

IV. Experiments

Input Image Collection. Following BiasPainter [16], we focus the scope of our experiments by considering only 2 genders (male and female), 3 races (East Asian, White, Black), and 5 age categories (20-29, 30-39, 40-49, 50-59, 60-69). Our seed images are selected from the FairFace dataset [24], which includes the ground truth labels for gender, race, and age. We select 3 images from each combination of gender, race, and age, leading to a total of 90 images.

Neutral Prompt List Collection. We compile a list of 90 neutral prompts using 42 professions and 48 personalities, with the following templates: '*a person who is a/an [profession]*' and '*a person who is [personality]*' [16].

Image Generation. The input images and prompts are then processed through both the original image editing model and GENIE. To provide debiasing guidance, the debiasing terms must be carefully selected to steer the model in an intended direction, avoiding any ambiguous guidance. In our experiments, debiasing terms are defined using the ground truth labels from FairFace [24]. Alternatively, a classifier can be used for attribute extraction. For gender bias, 'male' or 'female' is used based on the individual's gender in the image. For age bias, age is quantised to five possible values $(20, 30, \ldots 60)$ and is chosen depending on the age of the individual. Lastly, for race bias, the individual's race is chosen as the debiasing term. Four widely used editing models are selected for our experiments: IP2P, SD1.5, SD2.1, SDXL. In total, we will have 90 images $\times 90$ prompts $= 8100$ sets of (input image, prompt, generated image before debiasing, generated image after debiasing) to evaluate the debiasing effects of our method on each model.

Bias Quantification. To measure gender, race, and age biases effectively, we adopt the bias scores proposed in [16]. The scores will be computed twice, once with the generated images prior to debiasing and again after debiasing.

***Gender Bias Score*.** The gender of the generated image can be identified using a gender classifier, for which we have selected DeepFace [25]. The gender bias score is calculated by comparing the gender of individuals in both the input and generated images. A score of 0 is assigned if the gender remains the same. A change from male to female results in a score of 1, while female to male receives a score of -1.

Age Bias Score. This score is calculated by comparing the age difference between the input and generated images. To obtain the age labels for the generated images, an age classifier from Hugging Face [26] is used. The age difference $(output\_age - input\_age)$ is divided by 20 so the score ranges from -2 to $+2$. The score is designed such that a score of a ± 1 would indicate a significant age generation change (e.g. middle age to old age) due to age bias.

$$\text{Age\_Bias\_Score} = \frac{output\_age - input\_age}{age\_threshold} \tag{6}$$

Race Bias Score. The race bias score is computed by comparing the average grayscale values of skin tones between the input and generated images. Before comparing, filtering is applied to normalise the lighting across images. A score of 0 indicates no change in skin tone in the generated image, while a positive score indicates a lighter skin tone and a negative score indicates a darker skin tone. We divide the grayscale differences by 20. While the score can go beyond ± 1, similarly, the score is designed such that a score of ± 1 indicates a clear change in skin tone [16].

$$\text{Race\_Bias\_Score} = \frac{output\_grayscale - input\_grayscale}{race\_threshold} \tag{7}$$

Word Bias Score. The word bias score is calculated to determine the biases associated with each prompt word in each model. It can be calculated by finding the sum of the gender/race/age bias scores for each prompt divided by the total number of images N. X represents gender, age, or race.

$$\text{Word\_X\_Score} = \frac{\sum_{i=0}^{N} \text{X\_Bias\_Score}_i}{N} \tag{8}$$

Model Bias Score. The model bias score can be calculated by summing the absolute values of all word biases and dividing it by the total number of prompt words, denoted as M.

$$\text{Model\_X\_Score} = \frac{\sum_{i=0}^{M} |\text{Word\_X\_Score}_i|}{M} \tag{9}$$

V. Results

A. Quantitative Results

Table I shows bias scores for gender, age, race across different models before (Ori) and after applying GENIE (Ours).

TABLE I: Model Bias Scores Before and After Debiasing.

		Gender		Age		Race	
Model	**Domain**	**Ori**	**Ours**	**Ori**	**Ours**	**Ori**	**Ours**
IP2P	Profession	0.26	0.08↓	0.44	0.13↓	0.40	0.23↓
	Personality	0.17	0.15↓	0.10	0.07↓	0.10	0.07↓
SD1.5	Profession	0.27	0.15↓	0.79	0.54↓	0.13	0.05↓
	Personality	0.16	0.16	0.63	0.63	0.04	0.03↓
SD2.1	Profession	0.14	0.11↓	0.47	0.51	0.03	0.03
	Personality	0.12	0.10↓	0.45	0.45	0.03	0.03
SDXL	Profession	0.14	0.09↓	0.65	0.65	0.03	0.02↓
	Personality	0.05	0.06	0.73	0.74	0.03	0.02↓

For gender bias, substantial reductions are observed across all models. Notably, gender bias in the profession domain is decreased from 0.26 to 0.08 in IP2P, representing a 69% reduction while SD1.5 shows a 44% decrease from 0.27 to 0.15. For age bias, our model shows mixed results. Our method can greatly reduce age biases in the profession domain in both IP2P and SD1.5. However, SD2.1 and SDXL show no decrease in age bias; and we posit that this is likely due to the unnatural edits by those two models, where age-related features such as wrinkles are smoothed out, causing the models to fail to preserve age features. This is also evident in Figure 2's mechanic output (produced by SD2.1), where although the mechanic's age seems preserved, the skin appears noticeably smoothed. Lastly, our approach also performs well in mitigating race biases. In the profession domain, race bias in IP2P and SD1.5 is reduced by close to 50%.

Fig. 2: Illustrative comparison of debiasing effects using GENIE vs the original models using a text prompt of 'A person who looks like a [profession]'.

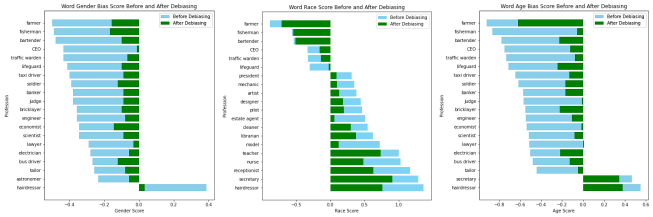

Fig. 3: Word Scores across top-20 most biased professions in IP2P: Gender (left), Race (centre), Age (right). Blue bars represent pre-debiasing scores, green bars post-debiasing.

Figure 3 visualises the word scores across IP2P's top-20 most biased professions. Before debiasing, the model exhibits strong biases. As shown by the blue bars, many professions are associated with being male (negative gender scores), young (negative age scores), and fair skinned (positive race scores). After debiasing, the bias scores (green) are significantly reduced, showing that GENIE effectively mitigates biases.

B. Qualitative Results

We present results from both the original editing model and GENIE in Figure 2 to observe the debiasing effects of our model. In the original models, the biases are manifested in the change of female to male for 'mechanic' and the lightening of skin tone for 'CEO'. In contrast, GENIE can preserve these attributes. For instance, the skin tone of 'CEO' is better preserved and the person's gender in 'mechanic' is maintained. Furthermore, our approach succeeds in performing occupation-related edits while preserving the important attributes, showing that our method can effectively mitigate biases.

VI. Conclusion

GENIE provides a guidance-based debiasing methodology to mitigate biases in text-to-image editing models. Experiments on 4 popular open-sourced models demonstrate the effectiveness of GENIE, reducing bias, on average, by 31% in gender, 15% in age, and 39% in race.

REFERENCES

[1] OpenAI. (2024) Openai. [Online]. Available: https://openai.com/

[2] ——. (2024) Dall-e3. [Online]. Available: https://openai.com/dall-e-3

[3] ——. (2024) Chatgpt. [Online]. Available: https://openai.com/chatgpt

[4] ——. (2023) Dalle3 is now available in chatgpt plus and enterprise. [Online]. Available: https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise

[5] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3043–3054.

[6] R. Naik and B. Nushi, "Social biases through the text-to-image generation lens," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 786–808.

[7] S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite, "Stable bias: Evaluating societal representations in diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] A. Birhane, S. Han, V. Boddeti, S. Luccioni *et al.*, "Into the laion's den: Investigating hate in multimodal datasets," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[9] M. Kay, C. Matuszek, and S. A. Munson, "Unequal representation and gender stereotypes in image search results for occupations," in *Proceedings of the 33rd annual acm conference on human factors in computing systems*, 2015, pp. 3819–3828.

[10] Y. Wu, Y. Nakashima, and N. Garcia, "Gender bias evaluation in text-to-image generation: A survey," *arXiv preprint arXiv:2408.11358*, 2024.

[11] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente *et al.*, "Holistic evaluation of text-to-image models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan, "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1493–1504.

[13] Y. Wan, A. Subramonian, A. Ovalle, Z. Lin, A. Suvarna, C. Chance, H. Bansal, R. Pattichis, and K.-W. Chang, "Survey of bias in text-to-image generation: Definition, evaluation, and mitigation," *arXiv preprint arXiv:2404.01030*, 2024.

[14] M. D'Incà, E. Peruzzo, M. Mancini, D. Xu, V. Goel, X. Xu, Z. Wang, H. Shi, and N. Sebe, "Openbias: Open-set bias detection in text-to-image generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 225–12 235.

[15] M. Cheong, E. Abedin, M. Ferreira, R. Reimann, S. Chalson, P. Robinson, J. Byrne, L. Ruppanner, M. Alfano, and C. Klein, "Investigating gender and racial biases in dall-e mini images," *ACM Journal on Responsible Computing*, 2023.

[16] W. Wang, H. Bai, J.-t. Huang, Y. Wan, Y. Yuan, H. Qiu, N. Peng, and M. R. Lyu, "New job, new gender? measuring the social bias in image generation models," *arXiv preprint arXiv:2401.00763*, 2024.

[17] M. M. Tanjim, K. K. Singh, K. Kafle, R. Sinha, and G. W. Cottrell, "Discovering and mitigating biases in clip-based image editing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2984–2993.

[18] F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, and K. Kersting, "Fair diffusion: Instructing text-to-image generation models on fairness," *arXiv preprint arXiv:2302.10893*, 2023.

[19] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 522–22 531.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[21] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[22] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[23] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.

[24] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1548–1558.

[25] S. I. Serengil and A. Ozpinar, "Hyperextended lightface: A facial attribute analysis framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/9659697

[26] Nate Raw, "vit-age-classifier (revision 461a4c4)," 2023. [Online]. Available: https://huggingface.co/nateraw/vit-age-classifier