Latent Directions: A Simple Pathway to Bias Mitigation in Generative AI

Carolina López Olmos[1,2] Alexandros Neophytou[2] Sunando Sengupta[2] Dim P. Papadopoulos[1]
[1] Technical University of Denmark [2] Microsoft
{clopezolmos, alexandros.neophytou, sunando.sengupta}@microsoft.com, dimp@dtu.dk

Abstract

Mitigating biases in generative AI and, particularly in text-to-image models, is of high importance given their growing implications in society. The biased datasets used for training pose challenges in ensuring the responsible development of these models, and mitigation through hard prompting or embedding alteration, are the most common present solutions. Our work introduces a novel approach to achieve diverse and inclusive synthetic images by learning a direction in the latent space and solely modifying the initial Gaussian noise provided for the diffusion process. Maintaining a neutral prompt and untouched embeddings, this approach successfully adapts to diverse debiasing scenarios, such as geographical biases. Moreover, our work proves it is possible to linearly combine these learned latent directions to introduce new mitigations, and if desired, integrate it with text embedding adjustments. Furthermore, text-to-image models lack transparency for assessing bias in outputs, unless visually inspected. Thus, we provide a tool to empower developers to select their desired concepts to mitigate. The project page with code is available online[1].

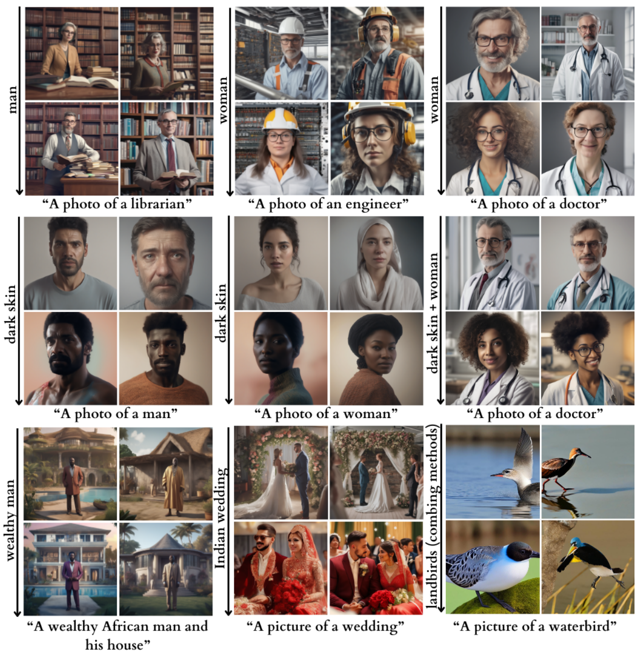

Figure 1. **Debiasing diverse concepts.** Generations observed upon the application of our approach in several scenarios.

1. Introduction

Text-to-image models have enabled the possibility of generating personalized images with the content described by words, transforming industries, and even, our thoughts. Given these models' impact on our lives, it is key to guarantee they are developed responsibly, battling stereotypes [16, 25], lack of diversity, and inherited biases, but remaining truthful [22]. Moreover, if this biased generated data is used for training future models, these biases will persist or be amplified in the new models themselves.

From gender to race, to poor geographic representation, biases can be found everywhere in generative models. Leonardo N. and Dina B. [16] have found a concerning pattern in Stable Diffusion v1.5 where low-paying jobs are dominated by women and darker-skinned individuals. While most of the work focuses on social and racial biases [4, 27], Basu *et al.* [2] conduct a user study validating geographical biases in models like DALLE [21] and Stable Diffusion [23], finding underrepresented 25 out of 27 countries. Cultural biases are also found when using homoglyphs in text-to-image synthesis [15]. Efforts to understand biases in vision-language models lead to the development of automated tools [6, 14], the use of gender estimation [13], face and skin tone detection [5, 9], and the evaluation of geographical representativeness using CLIP-based similarity and k-nearest neighbor models [2]. Mitigation efforts include prompt interventions [1], the development of more inclusive datasets featuring images reflecting diverse geographic and socioeconomic contexts [11, 20], and alterations in prompt embedding strategies [7, 28]. Chuang *et al.* [7] propose a method to mitigate bias by maximizing the similarity between biased and non-biased prompts. They

[1] https://latent-debiasing-directions.compute.dtu.dk/

construct a projection matrix to eliminate biased directions from text embeddings before inputting them into the model. Using a similar approach, but employing tokens of available image datasets, ITI-Gen [28] learns a set of prompt embeddings to append to the initial prompt.

We first present a tool to enhance developers' visibility, given that we believe understanding the relationship between concepts, and the reason for certain attributes appearing in generations, is key to mitigating them. Secondly, we propose a straightforward novel approach for bias mitigation, linearly separating the latents, noisy information tensors of two different prompts, learning the transformation for debiasing in the latent space of the diffusion model.

We apply this transformation, named *latent direction*, at a specific weight, linearly combining it with the initial Gaussian noise. The results in a series of diverse experiments prove it successfully works to debias, without the need for prompt alteration. Our approach remains simple, while effective and adaptable to varied scenarios. It allows the synergy of different latent directions, and it is flexible to be used in combination with an approach modifying the prompt embeddings if desired. Through our experiments, we focus on demonstrating the impact of our method for the maximum debiasing transition. However, fair distributions[2] can be obtained when applying the learned latent direction to only a determined percentage of the generations.

2. A Tool for Bias Understanding

Our tool for bias understanding targets two key points: comprehending the connections between embeddings and generations, and detecting the social characteristics and objects presented in the image. Theoretically, the closer the relation between attribute and concept in the semantic space, the more prone these attributes are to appear in the generated images. We explain the semantic relationship between attributes and concepts by computing the cosine similarity of their embeddings, and comprehending the innate biases within the employed text and vision encoders. In addition, we reveal the visual components of the generated images, using CLIP [19] as a zero-shot classifier for gender and race, and Kosmos-2 [18] as a Multimodal Large Language Model (MLLM) for perceiving object descriptions from the visual output seen in the images. With this, we present the frequency of objects and social characteristics in the generations, validate if the embedding associations correspond to the visualized content, and provide an understanding of the results without seeing the images. For instance, Fig. 2 informs us that our generations are debiased, with men in suits in front of their houses. However, it presents the innate biases of CLIP's text and vision encoders in Stable Diffusion

[2]Distribution of the generations selected by the user with ethical, truthful, and responsible considerations in mind.

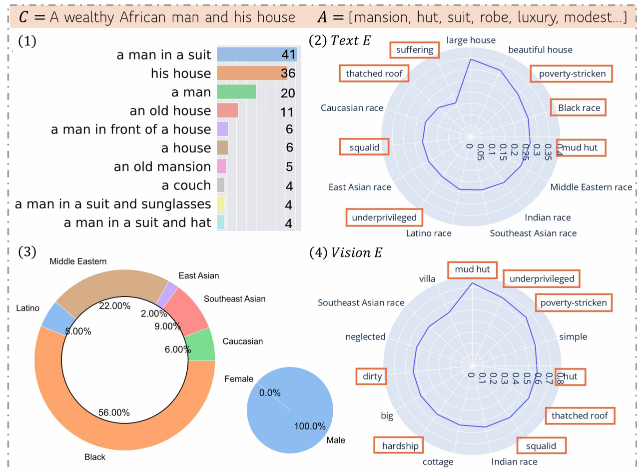

Figure 2. **Example of automated tool output.** Analysis of 100 generations of the concept C across 50 attributes A. (1) Frequency count of visual components across the images. (2, 4) Top 15 attributes exhibiting the highest cosine similarity (C, A) across text and vision encoders. (3) Gender and race detections.

2.1 where despite using the prompt *"A **wealthy** African man and his house"* the highest embedding similarities belong to attributes such as *poverty-stricken* or *underprivileged*.

3. Our Proposed Method for Bias Mitigation

Training: Finding the Latent Direction Our approach (Fig. 3) proposes a fundamentally different transformation of the diffusion process's input, learning the latent direction d_Z, from the Gaussian latents at denoised steps, to condition the initial noisy information tensor fed into the diffusion process $z_T \sim \mathcal{N}(0, I)$. Given a pre-trained latent diffusion model (LDM) and a neutral prompt P_1 (*e.g., "a photo of a man, in color, realistic, 8k"*), we aim to obtain debiased generations in the absence of prompt modifications or embeddings alterations. We propose a 'target' prompt P_2 (*e.g., "a photo of a black man, in color, realistic, 8k"*) and sample N number of images, for both P_1 and P_2, to construct the training dataset. The diffusion process for each of the prompts starts from an initial noisy latent z_T, which is denoised over k steps, finally reaching the ultimate latent z, fed into the decoder D to generate the synthetic image $\tilde{x}$.

While generating the N images, we save each image's latents at chosen denoising steps $L = (L_0, \cdots, L_k)$, representing $(z_T, \cdots, z_{T-i}$ for $i \in \{0, 1, 2, \ldots, k\}$), building a dataset of noisy information tensors. Note, L_0 corresponds to the initial Gaussian latent z_T, while L_k is z. Once all chosen latents are saved for both prompts, we select one denoising step i to obtain d_Z with those specific latents. For instance, we could decide to train with L_{10}, the latents saved for the N images at step 10 (z_{T-10}).

The model. We use a support vector machine (SVM) [26]

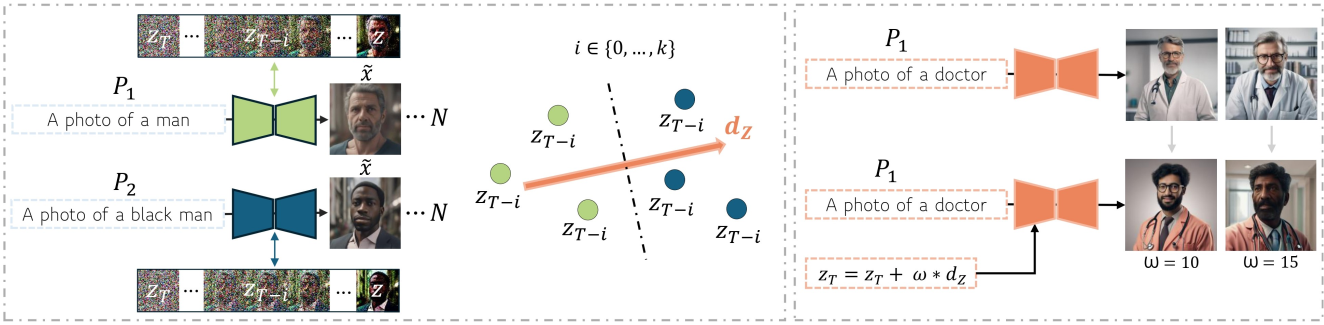

Figure 3. **Summary of our training (left) and inference (right) approach.** We use P_1 and P_2 to generate N images $\tilde{x}$. With their latents, chosen at step i, we train a SVM to learn d_Z. We debias the neutral prompt P_1, applying d_Z to the random initial latent $z_T \sim N(\mu, \sigma^2)$ at a specific ω weight, shifting the generations towards debiased samples with the attributes learned through the latent direction.

to linearly separate the latents across our labeled dataset of N samples for each prompt. The classifier uses a linear kernel and provides the d_Z, the so-called *latent direction*, we utilize for debiasing.

Inference: Applying the Latent Direction. To obtain debiased generations, the LDM, in our case Stable Diffusion [23], uses for inference only P_1, the neutral prompt. This prompt is fed into CLIP's text encoder E forming the first input. As the second one, instead of using an initial Gaussian random information tensor for denoising, we transform this latent by applying the learned latent direction d_Z, following equation 1.

$$z_T = z_T + \omega \cdot d_Z \tag{1}$$

Where $z_T \sim N(\mu, \sigma^2)$ and ω is the weight parameter at which the latent direction is applied. The higher the ω, the higher the strength of the debiasing impact.

The optimal configuration. Optimal debiasing results are found when selecting the most favorable latent $L = (L_0, \cdots, L_k)$ and weight configuration ω. Thus, we propose two approaches to automatically find it without having to visually explore all possibilities. The first one is to use the *clean-fid* [17] library to compute the similarity between the distribution of a small subset of generated images with a particular configuration, and the distribution of known debiased images, and the second one is to leverage CLIP [19] as a zero-shot classifier, selecting the configuration with a high classification of the desired debiased class.

4. Experimental Results

We validate our work using Stable Diffusion XL [23], with 50 denoising steps, applying different latent directions d_Z in a series of experiments to understand its impact. Successful results are obtained in diverse mitigations. We present four different debiasing scenarios following the settings of previous papers [7, 10, 12, 28], addressing social group biases, cultural and geographical biases, and the Waterbird [24] benchmark for evaluating spurious correlations. Fig. 1 summarizes the experimental results obtained.

Quantitative Metrics. We leverage the Statistical Parity Difference (SPD) [8] to evaluate our debiasing method in the generated image datasets. We use CLIP for attribute prediction and measure the absolute difference in the proportions of desired attributes between the original biased dataset, generated with the plain Stable Diffusion model, and the debiased dataset. A value close to zero indicates minimal debiasing impact, while a value of one signifies successful debiasing with the desired attribute present in all generations.

Gender debiasing in professions. We learn d_Z by defining $N = 50$, $P_1 =$*"a photo of a man, in color, realistic, 8k"* and $P_2 =$*"a photo of a woman, in color, realistic, 8k"*, selecting L_{25} and $\omega = 10$. We apply the *woman* latent direction to the neutral prompts *"a photo of a [profession], in color, realistic, 8k"* and observe a positive shift from 0% to 52%, in 100 generations for the case of "doctor". Other professions known to be extremely biased, such as firefighter, engineer, or librarian have shown a slightly improved impact, with shifts of 8%, 3%, and 2%, respectively. We believe major debiasing can be achieved with these professions upon finding the optimal d_Z.

Skin tone debiasing. We explore the transition of skin tones in generated images. For it, we create four training datasets, where $N = 50$, with $P_1 =$*"a photo of a [man/woman], in color, realistic, 8k"* and $P_2 =$*"a photo of a black [man/woman], in color, realistic, 8k."*. With the proposed automated method for configuration selection we settle on training with the latents at step 10 (L_{10}), at a weight $\omega = 15$. The results shift 100 generations using the neutral prompt $P_1 =$*"a photo of a man, in color, realistic, 8k"* to contain a 95% of black men images from the original 8%. Similarly, with $P_1 =$*"a photo of a woman, in color, realistic, 8k"* and the application of the dark-skin d_Z at L_{25}, $\omega = 14$ yields a 79% of black women from an initial 1%.

Waterbird debiasing. In this experiment we evaluate the

d_Z	Skin Tone		Gender			Landbird	Indian	Wealth
P_1	*Man*	*Woman*	*Doctor*	*Firefighter*	*Cleaner*	*Waterbird*	*Wedding*	*African man*
(SD XL, ours)	0.87	0.78	**0.52**	**0.08**	0.09	-	0.33	0.28
(SD 2.1, PD [7])	0.91	0.90	0.14	0.06	0.01	0.29	0.79	0.47
(SD 2.1, ours + PD [7])	**0.94**	**1.00**	0.29	0.04	**0.22**	**0.68**	**1.00**	**0.96**

Table 1. **Quantitative results across 100 generations.** SPD in the presence[3] of the desired attributes: dark skin tone, female gender for the case of *doctor, firefighter* and male gender for *cleaner*, land environments for waterbirds, Indian wedding attributes and wealthier looking houses avoiding thatched roofs and mud huts. We learn the latent directions with SD 2.1 for the results seen in the last row.

impact of the combination of methods, manipulating both the prompt's embeddings [7] and the initial Gaussian noise, using Stable Diffusion 2.1. We aim to generate waterbirds in land environments with the prompt P_1 = *"A picture of a waterbird"*. We replicate their setup and apply our learned latent direction (P_1 = *"A picture of a waterbird"*, P_2 = *"A picture of a landbird"*, $\omega = 10$, L_{10}). The results across 100 generations yield exceptional results with 78% of generations displaying waterbirds in terrestrial habitats, 2% in aquatic landscapes, and 20% showing bird portraits.

Geographical representativeness. It is hard to obtain balanced geographical representations of *"a picture of a wedding, in color, realistic, 8k*", given this neutral prompt is normally biased towards representations of Western weddings. In an attempt to shift the distribution towards Indian weddings, we define P_1 =*"a picture of a wedding, in color, realistic, 8k"* and P_2 =*"a picture of a wedding in India, in color, realistic, 8k"*, learning d_Z using L_{30} and applying it with $\omega = 35$ we see an increment of 33% in CLIP's classification. Inspired by Bianchi *et al.* [3] we debias 38% of the thatched roofs observed when using P_1=*"A wealthy African man and his house"*, utilizing P_2=*"A wealthy man and his house"* and applying the learned *wealthy man* direction d_Z at L_{10} and $\omega = 15$.

Comparison with PD. In Tab. 1, we present the quantitative results of our study. The comparison with the Prompt Debiasing (PD) method [7] is challenging, due to the utilization of distinct models and the diverse biases in them, e.g., for P_1 =*"A wealthy African man and his house"* SD XL presents generations of mansions with thatched roofs, whereas SD 2.1 shows mud huts. We choose to use SD XL given the enhanced quality of the model facilitates the learning of d_Z and minimizes the inconveniences of elaborated hard prompting to obtain quality images with SD 2.1. The outcomes evaluated through our experiments demonstrate the potential of latent directions to obtain competitive debiased generations despite maintaining neutral embeddings.

4.1. Relevant Insights and Learnings

The integration of prompt debiasing and latent directions surpasses the efficacy of the former used individually (Tab. 1). Regarding our approach, the results in Fig. 4 confirm the choice of weight ω has a higher impact on the debiasing than the choice of training latent L. Moreover, higher latent directions require lower weights to achieve the debiased results, given more structured noise is found at the higher debiasing steps. However, as we move in d_Z there is a limit to how far we go with ω, given an extremely high weight leads to distorted generations, out of the distribution. Lastly, it is possible to linearly combine latent directions following $z_T = z_T + \sum_{i=1}^{\infty} \omega_i \cdot d_{Zi}$. For instance, by applying the *woman* [L_{25} $\omega 10$] and *dark-skin* [L_{10} $\omega 10$] latent directions to the Gaussian noise of the neutral prompt P_1 =*"a photo of a doctor, in color, realistic, 8k"* we achieve generations of dark-skinned female doctors (Fig. 1).

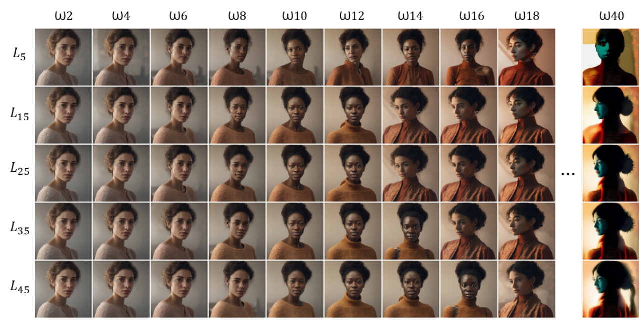

Figure 4. **Comparison of results with d_Z trained at different latents L and applied at different weights ω.** Generations of the same woman in its transition to dark skin.

5. Conclusion

After proposing a tool for uncovering and quantifying the present bias in text-to-image models, a novel method is proposed for mitigation. By learning and applying latent directions d_Z we demonstrate it is possible to alter the diverse complex biased relations, such as those in cultural events, while maintaining unaltered neutral prompt embeddings. **Future work** encourages the exploration of more advanced classifiers to find the optimal d_Z.

[3]Presence of classes through CLIP's classification: ["A picture of a black [man/woman]", "A picture of a white [man/woman]"], ["A picture of a woman", "A picture of a man"], ["A picture of a Western wedding", "A picture of an Indian wedding"]. A user study is used to evaluate the complex generations (*Wedding*, *African man*) given classification with defined classes in these cases does not match reality.

References

[1] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions?, 2022. 1

[2] Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models, 2023. 1

[3] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2023. 4

[4] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016. 1

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. 1

[6] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models, 2023. 1

[7] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts, 2023. 1, 3, 4

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011. 3

[9] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation, 2022. 1

[10] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness, 2023. 3

[11] William Gaviria Rojas, Sudnya Diamos, Keertan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Advances in Neural Information Processing Systems*, pages 12979–12990. Curran Associates, Inc., 2022. 1

[12] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses, 2021. 3

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 1

[14] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models, 2023. 1

[15] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation, 2013. 1

[16] Leonardo Nicoletti and Dina Bass. Humans are biased. generative ai is even worse. 2023. Accessed: 2023-09-10. 1

[17] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation, 2022. 3

[18] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. 2

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 3

[20] Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B. Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Geode: a geographically diverse evaluation dataset for object recognition, 2023. 1

[21] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 1

[22] Adi Robertson. Google's ai model gemini criticized for generating historically inaccurate images. *The Verge*, 2024. 1

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 3

[24] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. 3

[25] Irene Solbes-Canales, Susana Valverde-Montesino, and Pablo Herranz-Hernández. Socialization of gender stereotypes related to attributes and professions among young spanish school-aged children. *Frontiers in Psychology*, 11, 2020. 1

[26] V. N. Vapnik and A. Ya. Chervonenkis. *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, pages 11–30. Springer International Publishing, 2015. 2

[27] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks, 2018. 1

[28] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Iti-gen: Inclusive text-to-image generation, 2023. 1, 2, 3