DebiasPI: Inference-time Debiasing by Prompt Iteration of a Text-to-Image Generative Model

Sarah Bonna, Yu-Cheng Huang, Ekaterina Novozhilova, Sejin Paik, Zhengyang Shan, Michelle Yilin Feng, Ge Gao, Yonish Tayal, Rushil Kulkarni, Jialin Yu, Nupur Divekar, Deepti Ghadiyaram, Derry Wijaya, and Margrit Betke *

Boston University, Boston MA 02215, USA

Abstract. Ethical intervention prompting has emerged as a tool to counter demographic biases of text-to-image generative AI models. Existing solutions either require to retrain the model or struggle to generate images that reflect desired distributions on gender and race. We propose an inference-time process called DebiasPI for Debiasing-by-Prompt-Iteration that provides prompt intervention by enabling the user to control the distributions of individuals' demographic attributes in image generation. DebiasPI keeps track of which attributes have been generated either by probing the internal state of the model or by using external attribute classifiers. Its control loop guides the text-to-image model to select not yet sufficiently represented attributes, With DebiasPI, we were able to create images with equal representations of race and gender that visualize challenging concepts of news headlines. We also experimented with the attributes age, body type, profession, and skin tone, and measured how attributes change when our intervention prompt targets the distribution of an unrelated attribute type. We found, for example, if the text-to-image model is asked to balance racial representation, gender representation improves but the skin tone becomes less diverse. Attempts to cover a wide range of skin colors with various intervention prompts showed that the model struggles to generate the palest skin tones. We conducted various ablation studies, in which we removed DebiasPI's attribute control, that reveal the model's propensity to generate young, male characters. It sometimes visualized career success by generating two-panel images with a pre-success dark-skinned person becoming light-skinned with success, or switching gender from pre-success female to post-success male, thus further motivating ethical intervention prompting with DebiasPI.

Keywords: Generative AI · Racial and gender bias · Debiasing

1 Introduction

Generative AI models have made their mark in journalism, with AI-generated images that accompany news articles [3, 27], sometimes even without disclosing

* Bonna and Huang are co-first authors with equally important contributions. Corresponding authors {sbonna,ychuang2,betke}@bu.edu.

Fig. 1: AI model visualizing the news headline: "From School Janitor to Esteemed School Superintendent" without (left) and with (middle, right) prompt intervention. The two-panel image on the left is supposed to show the same person at different stages of their life, but the janitor is depicted as Black and the superintendent as White.

the use of AI [24]. Given the increased exposure of the public to AI-generated news-accompanying images, it is concerning that analysis of AI-generated images has revealed levels of racial and gender biases [6], for example, sexualized images of women of color [12]. Such images perpetuate and amplify stereotypes and could spread on the internet in connection with digital news. Recent studies indicate that biases of text-to-image AI models extend across various dimensions of generated content, including skin tones, gender, and attire [6, 10, 23]. The research question has arisen: To what extent can ethical interventions via prompting influence generative text-to-image AI models to produce outputs that ensure diverse representations of people?

Our research addresses this question by building on the idea of "prompting with ethical intervention:" Recent work [11] designed a procedure for training text-to-image AI models that changes the demographic attribute of a person in a prompt according to a desired input distribution of the attribute. Other work experimented with prompts such as "a person who works as a nurse" to diagnose social bias of the model [9] and prompts with ethical interventions, e.g., "a photo of a bride from diverse cultures," to mitigate the social bias of the model [2]. While these prior works provided an important proof-of-concept of the idea of "prompting with ethical intervention," they require training of the generative models, which was accomplished for relatively small text-to-image generative models (DALL-E$^{\text{Small}}$ [26], minDALL-E [15], and Stable Diffusion [20]).

Our study addresses the task from the perspective of a newsroom editor who cannot retrain or finetune a text-to-image model and would like to make a selection from a set of images created by a commercial tool. We selected DALL-E 3 [5] as the text-to-image model in our experiments. As part of our methodology, we generated demographics-neutral news headlines, specifically about human-interest career success stories, including "rags-to-riches stories," and asked DALL-E 3 to interpret these headlines visually. Our motivation for using the success story theme was the expectation that if generative AI was used in the news, it should be able to provide inspiring images about a diverse set of people and thus try to influence societal narratives in a positive way. An example of a headline and generated images are shown in Fig. 1.

We introduce an inference-time process, called Debiasing by Prompt Iteration (DebiasPI), which is designed to support a user of a given text-to-image model, for example, a news room editor, in obtaining images of people with demographic attributes that follow a desired input distribution. DebiasPI keeps track of the attributes of people in the generated images, guiding the text-to-image model to select specific attributes. DebiasPI has two mechanisms to do this: it can either use the internal believe of the model in the attribute it has generated, or it can use external classifiers to evaluate the attribute in the generated image. We also provide tools for comparing the desired and obtained attribute distributions, which inform users on the state of the debiasing process and whether it has converged.

In addition to automated attribute evaluation, we also provide a human-based evaluation process, using *quantitative content analysis* (QCA), a research methodology employed by journalism scholars to evaluate communication artifacts [4, 18]. We developed a codebook, the data collection instrument used in content analysis, to guide human annotators in labeling the perceived attributes in the AI-generated images, such as race, gender, skin tone, body type, and age. Following best practices in QCA, we pretested the codebook for intercoder-reliability before annotating the images.
In summary, the contributions of this work are:

- DebiasPI, a debiasing-by-prompt-iteration inference-time process that enables ethical prompt intervention by controlling distributions of individuals' demographic attributes in image generation;
- A codebook for manual annotation of skin tone, race, gender, body type, and age of people in AI-generated images, as well as recommendations for tools to evaluate generated attributes and their distributions;
- Textual and visual generative datasets concerning "rags-to-riches" news stories, which can serve benchmark comparisons by others in future work.
- Experimental results of DebiasPI, prompting with and without ethical interventions.

The code for DebiasPI and its analysis tools, the textual and visual generative datasets of our various experiments, the annotations, and the codebook are available at `http://www.cs.bu.edu/faculty/betke/research/DebiasPI`.

2 Related Work

Bias, stereotypes, and representational harm have been identified as areas of impact that generative AI may have on society [22]. Bias can be introduced at various stages of the machine learning pipeline - the model used, compression techniques, and many other factors can "amplify harm on underrepresented protected attributes" [22]. Besides these factors, the characteristics of the researchers and the developer organizations can introduce biases too, such as the structure, demographics, and geographic location of the team. Biases caused by a lack of representation while training and developing the generative AI system

can marginalize already marginalized groups even more when such AI systems are deployed.

Recent studies have highlighted the presence of biases in various dimensions within generative models. The study by Bianchi et al. [6] is an in-depth investigation of demographic stereotyping by image-generating AI models. The authors explored whether neutral wording about race, gender, ethnicity, and nationality in input prompts leads to the generation of harmful stereotypes in the output images. Their analysis found that harmful stereotypes are indeed generated by models, such as Stable Diffusion [20]. Bianchi et al. [6] describe a scenario where someone cleaning is depicted with stereotypically feminine characteristics; scenarios involving a poor person, a thug, or a person stealing yield faces with dark skin tones and stereotypically Black features. Prompting for an image with a terrorist results in brown faces with dark hair and beards, representing Middle Eastern men; prompting for an illegal person results in brown-skinned faces, meant to represent the perception of allegedly undocumented Latin American immigrants. The authors argue that all of these mentioned stereotypes are consistent with the American narrative perpetuated by the media and can incite violence and discrimination against these groups of people. The researchers have also found that AI models tend to amplify stereotypes specific to occupations.

Biases have been found in the portrayal of skin tones, genders, and specific garments in generated images [10]. Gender distributions vary across different professions, with a tendency to associate skirts primarily with women and suits, jackets, or ties with men. These biases may not be solely due to uneven attribute distribution in the training data but can also stem from a lack of detailed background context [23]. The absence of in-depth scene understanding and ignorance of the creator's intentions can lead to misrepresentations and incorrect patterns. Another study [12] found that Stable Diffusion [20] visualizes the term "personhood" typically as a Western male, while visualizations of women of color from India, Egypt, and Latin American countries were sexualized. Sun et al. [25] found that women are more likely to smile than men across occupational categories in DALL-E 2 images and their faces pitch downwards - a posture that may represent obedience or subordination.

In Large Language Models, biases manifest in responses to different demographic groups. For instance, GPT-4 has been observed to provide significantly different recommendations for diagnosis, assessment, and treatment for patients by only varying gender or race [28]. Similarly, ChatGPT shows certain inclinations in political orientation tests, yet it consistently avoids taking explicit stances on politically charged questions [21]. Addressing these biases involves multiple strategies beyond merely retraining models. One approach is "reinforcement learning from human feedback" [8], another one is ethical intervention prompting [29], a strategy we explore in our study.

Baseline Prompt:
Given a <text> about a person as input, your task is to generate a photograph that visualizes the person. Then output the generated image.
Prompt with Attribute List:
*Given a <text> about a person and an <**attribute list**>as inputs, your task is to select an attribute and generate a photograph that visualizes the person with the selected attribute. Then output the generated image and selected attribute.*
Prompt with Attribute Distribution:
*Given a <text> about a person and an <**attribute distribution**> as inputs, your task is to select an attribute according to the distribution and generate a photograph that visualizes the person with the selected attribute. Then output the generated image and selected attribute.*

Fig. 2: Three levels of prompting: The *Baseline Prompt* does not include any ethical intervention. The *Prompt with Attribute List* mentions attribute choices, while the *Prompt with Attribute Distribution* asks the model to choose attributes according to a desired distribution. A use case could be for the AI model to generate as many female as male entrepreneur pictures.

3 Method

3.1 Ethical Intervention Prompts and DebiasPI

We designed three types of prompts to evaluate the generation bias of text-to-image generative models, shown in Fig. 2. The baseline does not explicitly prompt the model to pay attention to any attributes, so it might emphasize certain traits and styles, according to its internal demographic bias. The *Prompt with Attribute Distribution* serves as a means to attempt to debias the generative model, using the proposed *Debiasing by Prompt Iteration (DebiasPI)* process visualized in Fig. 3.

The user of DebiasPI starts by setting a target distribution within the prompt. The distribution is defined by a list of attribute bins and the desired counts per bin. This list and the text (e.g., news headline) to be visualized are sent to the text-to-image model for image generation. The model response is parsed for chosen attributes, either using its internal belief or an external attribute classifier. The corresponding distribution bin is decreased by one, and once an attribute count reaches zero, the model is instructed to stop generating images with that attribute. Distribution statistics are collected throughout to determine the state of the debiasing process. The target distribution is reached once the counts of all bins are zero.

Initially during DebiasPI, the model utilizes its internal probability distribution for selecting attributes. As the allocated numbers are exhausted for certain attribute options and reach zero for the corresponding distribution bin, DebiasPI starts to adjust the generated attribute distribution to align more closely with our specified target distribution. This adjustment ensures that the final output adheres to the desired attribute proportions. By iteratively adjusting the prompt and the selection process, DebiasPI systematically reduces biases that may have

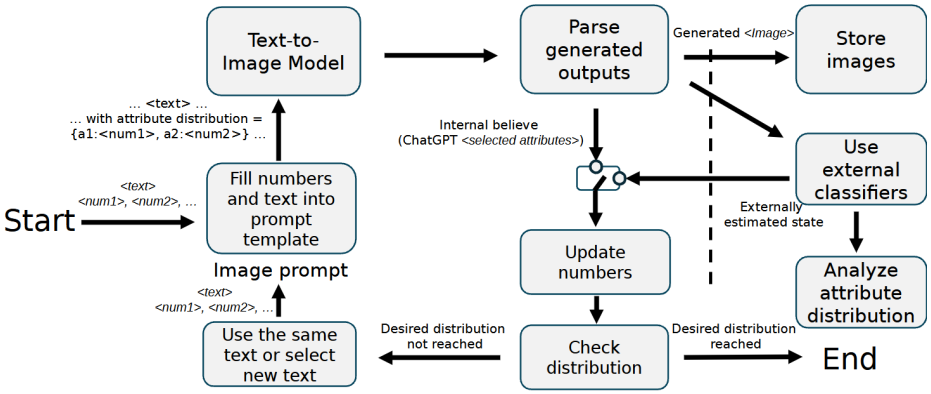

Fig. 3: Overview of the proposed Debiasing by Prompt Iteration (DebiasPI) process.

been present in the initial image generations of the model. For instance, if the internal probability distribution of the model favors certain attributes disproportionately, depleting those attributes to 0 forces the model to choose from the remaining attributes, thus redistributing the selection probabilities to align with our desired distribution.

For the DebiasPI process to be successful, it needs to keep track of the created attribute distribution at all times. It does this by parsing the generated images in one of two ways: (1) asking the text-to-image model directly what gender, race, etc., the depicted person belongs to, i.e., relying on the "internal belief" of the text-to-image model, or (2) using external classifiers to estimate gender, race, etc. of the person shown in the generated image, i.e., the "external belief." We describe the external classifiers that we deployed with DebiasPI with in Section 3.3.

If we aim for the generated distribution to match specific floating-point values of the desired attribute distributions, we must consider the potential for precision errors due to quantization effects with a limited number of generated images. To mitigate precision errors and enhance the accuracy of the generated distribution, it becomes essential to generate a larger number of images. By increasing the sample size, the attribute proportions can more accurately reflect the target distribution, thereby reducing the impact of any individual errors. Generally speaking, generating n images, where n is a power of 10, leads to a precision of n bits after the decimal point. This higher precision in larger sample sizes ensures a more accurate and representative final distribution, achieving a fairer and more balanced attribute distribution in the generated images.

To speed up the convergence of DebiasPI, the user can separate the generation process into several subgroups, each with the same distribution ratio. This way, the number of attributes can reach 0 earlier within each subgroup, allowing the system to start forcing the output to fit into the desired distribution with

less deviation. This subgrouping strategy accelerates the adjustment process, enabling the system to achieve the target distribution more efficiently.

3.2 Codebook for Manual Annotation of Attributes

The codebook we designed contains nine option-based questions about the race, gender, age, and occupation of the subject and various image characteristics (lighting, contrast, etc.), and an open-ended question about which (if any) stereotype the image might propagate, as perceived by the annotator.

Annotators were provided 10 swatches of skin tones from the Monk Skin Tone Scale [17] and given three options: Light (Types 1 to 3), Medium (Types 4 to 6), and Dark (Types 7 to 10). Examples of public figures around the world were shown below the swatch groups. The Monk Skin Tone Scale was chosen because it covered a large range of skin tones without overwhelming the annotators. To guide annotators in determining the race of a person in a generated image, an appendix was provided at the end of the codebook, with examples of a male and female public figure for each of nine races: Black, East Asian, Hispanic or Latino, Indigenous, Middle Eastern or North African, Native Hawaiian and Other Pacific Islander, South Asian, Southeast Asian, and White. The categorization into 9 races is derived from the race/ethnicity definitions by the U.S. Census Bureau. A map depicting the skin color of people in each region [7] was also provided as a reference. As humans are affected by the cross-race effect, it is hoped that these resources would help to reduce any confusion about the race of the subjects in the images.

The codebook offers three options for gender: Male, Female, and Unable to distinguish gender, four options for age: Children and adolescents (1–18), Young adulthood (19–35), Middle adulthood (36–64), and Seniors, and three options for body type: ectomorphs, mesomorphs, and endomorphs. Example images of public figures at the various age milestones (each decade) and graphics of body types are provided as references. Instead of asking the annotators to try to surmise the career or occupation of the person from the generated image, we instructed them to go back to the success story headline that was used to create the image and evaluate its perspective. In communication research, different perspectives are known as "frames", which, when used in news media, will influence the opinion of their readers in multiple ways. The codebook describes twelve frames that have been adapted from a list of occupations [1] (full list at `http://www.cs.bu.edu/faculty/betke/research/DebiasPI)`.

3.3 Methods to Evaluate Attributes and their Distributions

For automated analysis of skin tone of the main character's face in the generated image, we used the Facial Representation Learning in a Visual-Linguistic Manner (FaRL) model [30] to segment the largest area containing facial pixels. We then averaged the skin color within this area before quantizing it into the Monk Tone scale. For automated analysis of gender of a person in an AI generated image, we recommend the use of the Large Language and Vision Assistant (LLaVA) [16].

For estimating the age of a person in a generated image, we employed two Vision Transformer models [14, 19].

We use two measures to compare the distributions of desired and generated attributes: Given the desired distribution Q and the generated distribution P, the Jensen-Shannon Divergence (JS-Div)

$$\mathrm{JS}(P \parallel Q) = \frac{1}{2}\mathrm{KL}(P \parallel M) + \frac{1}{2}\mathrm{KL}(Q \parallel M) \tag{1}$$

is a symmetrized and smoothed version of the Kullback–Leibler divergence $\mathrm{KL}(P \parallel M) = \sum_i P(i) \log \frac{P(i)}{M(i)}$ with $M = 1/2(P + Q)$. Our second measure is the Earth Mover's Distance (EMD), which looks for a set of flows $\{f_{ij}\}$ between distribution bins i and j that minimizes the total cost of moving mass ("earth") from P to Q:

$$\mathrm{EMD}(P, Q) = \min_{\{f_{ij}\}} \frac{\sum_{i,j} f_{ij} d(i, j)}{\sum_{i,j} f_{ij}}, \tag{2}$$

where $d(i, j)$ is the ground distance between bins i and j. JS-Div and EMD are zero if $P = Q$. By employing these metrics, we can quantitatively assess whether and under which settings a text-to-image model can produce images with attribute distributions that align with our desired attribute distributions.

4 Experiments and Results

4.1 Headline Generation using GPT4-powered ChatGPT

We generated 200 headlines on human-interest success stories using GPT4-powered ChatGPT in four phases with 50 headlines per phase. Examples from each phase of the headline generation process were used as samples for the next phase. This iterative process was designed to vary the themes and sentence structures of the generated headlines. There were three primary headline patterns: 1) Transformation and overcoming adversity or hurdles; 2) Evolution of a role or growth within a profession; and 3) Journey from A to Z. We ensured that the generated headlines were demographic-neutral by specifically prompting ChatGPT to avoid including personal names, race, and gender. Examples of actual headlines were used to start the headline-creation process: "A Gen Z Success Story" (from the New York Times), "Teen entrepreneur shares sweet success story behind her multi-million-dollar lemonade business" (from Fox News). These two news outlets were selected to cover the two sides of the U.S. political spectrum, with The New York Times generally left-leaning and Fox News right-leaning. An example prompt is shown in Figure 4. The resulting headlines are available at `http://www.cs.bu.edu/faculty/betke/research/DebiasPI`.

4.2 Baseline Prompt Experiment: No Ethical Intervention

Iterating through the *Baseline Prompt* (Fig. 2) with <text>=headline 1,..., headline 200, we asked DALL-E 3 to create 200 images to visualize the generated

Headline Creating Prompt:
Generate 50 news headlines regarding the success stories of an individual, where success is defined as a spectacular transition from a low state of economic, academic, and or social being to a higher stage such that the transition is bordering on extraordinary and has a substantial impact on the real world, the person, or both. It's not just about the wildly successful success stories like Steve Jobs, but also those that demonstrate sincere grit, passion for upward mobility, and a deep desire to improve upon existing things and situations. Based on this definition of success, generate article/newspaper headlines of success stories. Do not include any names, gender, or race-identifying information in the headlines. Here are some sample headlines: The Phoenix Tale: Rising from the Ashes of Bankruptcy", "Food Cart Vendor Cooks Up a Culinary Empire" and "From Homeless to Harvard."

Fig. 4: ChatGPT prompt for creation of success story news headlines.

success story headlines, one image per headline text. We found that the model generated a male character in all but 3 images, i.e., 98.5% male and 1.5% female. The male figure was generally White (90%). These results were verified by human inspection and are aligned with the results of prior work [6], motivating ethical intervention prompting.

4.3 Experiment Yielding Two-Panel Images

Using the *Attribute List Prompt* (Fig. 2) with attribute lists of race or gender-&-race, we found that DALL-E 3 often created two-panel images (93 of 200 images), where the first panel shows a person before career success and the second panel after. The AI model fails to understand that the same person should be visualized and sometimes showed individuals of different race and/or gender, see Figs. 1(a) and 5. In cases when there is a difference in skin tone between the individuals shown in the two panels (14 images), our analysis showed that the change was always from a darker to a lighter skin tone, perpetuating the social bias that a person must be White or light-skinned to make career progress. Analysis of body types and occupations of the individuals revealed a preference of the model for the mesomorphs (athletic, solid, strong, not overweight or underweight). Endomorphs (lots of body fat or muscle) were rarely generated (8%). Images that depicted occupations in the areas of *Arts, Audio/Video Technology & Communications* and *Business Management, Administration & Finance* were predominately male (77%).

The "ground truth" of the attributes (prompted and unprompted) in this experiment were provided by four annotators. The Inter-Coder Reliability of each pair of annotators was calculated using the Cohen-Kappa score [13] and percent agreement based on an initial round of annotation on the first 40 images (10% of the total number of images available). The Cohen-Kappa scores ranged from 0.64 (Good) to 1.0 (Very Good). The percent agreement scores were then used to guide the allocation of annotator pairs to re-evaluate images for which the Cohen-Kappa score was only in the "Good" range. Using this process, we eventually

Fig. 5: Two-panel generated images obtained with attribute list prompts. The panel showing career success often showed a lighter-skinned person, usually a male.

achieved a Cohen-Kappa score greater than 0.8) [13], which is considered a robust Inter-Coder Reliability for QCA [18].

4.4 Ethical Intervention Experiments

Based on our experimental results with the *Attribute List Prompt* described in Sect. 4.3, in subsequent experiments with ethical intervention prompts, we adjusted the prompts shown in Fig. 2. We instructed the text-to-image model to generate a photograph-style image of a *single* person, facing forward in the generated image. This then enabled us to automate the annotation process, using the tools described in Section 3.3. We note that, given the list of 200 headlines, we could only prompt DALL-E 3 to process 5 headlines at a time, generating one image per headline, because it would hang when asked to generate more than five images at a time. This resulted resulted in 40 iterations of DebiasPI process.

Our experiments with the *Attribute Distribution Prompt* focused on the uniform distribution, asking the text-to-image model to produce outputs with equal representations of the attributes. As designed, the DebiasPI process results in outputs with a perfect balance of gender or race. The attribute type "skin tone" was more difficult to handle, as we show below.

To illustrate how DebiasPI iterates through its attribute selections and yields a balanced outcome, we report on an experiment with 50 images created with a 9-race uniform distribution prompt (Fig. 6). The plot shows how the selection of

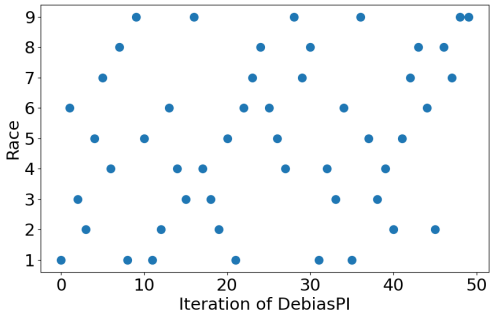

Fig. 6: Race choices made by DebiasPI during a 50-image generation process. Here, 6 Black individuals (race category 1) were frequently created by iteration 35, after the target is fulfilled the model does not generate any more black faces in subsequent iterations.

races is adjusted once the maximum (determined by the uniform distribution) for a specific race is reached.

Ablation Studies. We then asked the question: "What if we ask for a uniform distribution but handicap DebiasPI by not allowing it to keep track of the choices at each step?" The resulting "ablation studies" yield highly non-uniform distributions that show the text-to-image model's flawed interpretation of an equal representation of attributes. This outcome underscores the challenges in achieving truly balanced outputs of the text-to-image model without direct intervention. In the following, we describe the results of experiments with and without intervention on the attribute distributions of gender, race, age, and skin tone in detail.

Gender. We used the attribute list gender=[male, female]. Prompting the text-to-image model with this list of gender options yielded 71% males and 29% females. Prompting the model with an attribute list and desired uniform distribution, i.e., gender=[male, 50%, female 50%] for 200 images, yielded 56% males and 44% females, a moderately successful attempt at equal representation (in the absence of DebiasPI's explicit choice counting). Interestingly, we found that when the model was alerted to ethical interventions with respect to race but not gender, it improved the representation of gender in its outputs. Specifically, if race options were provided in the *Attribute List Prompt*, the resulting gender representation was 54% male and 46% female, and with a request for racial balancing, 52% male and 48% female.

Race. Next, we experimented with attribute lists race=[9 choices], and gender=[male, female], race=[9 choices]. The resulting race distributions are shown in Fig. 7 (again, in the absence of DebiasPI's explicit choice counting). Since we had to prompt DALL-E 3 to generate five images at a time, for each dataset creation, the model was called to generate images 40 times to cover the 200 headlines. DALL-E 3 only has 5 chances to cover 9 races in each of the 40 trials,

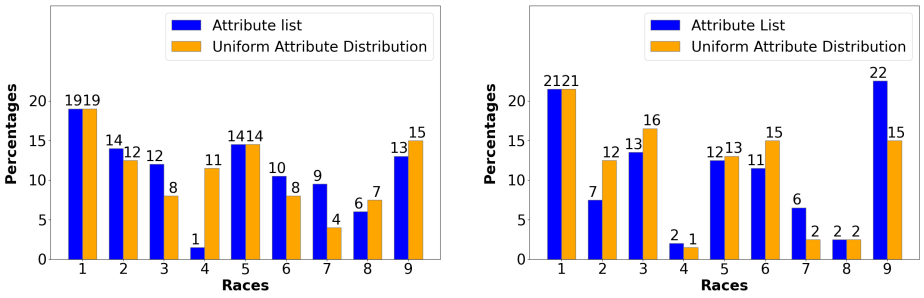

Fig. 7: Ablation study: Prompting with attribute lists only (blue) and with attribute distributions (orange) without DebiasPI's choice counting. Attribute lists are 9 race options (left) and 2 gender, 9 race options (right). The races are: Black, East Asian, Hispanic, Indigenous, Middle Eastern or North African, South Asian, Southeast Asian, Native Hawaiian and Other Pacific Islander, and White. The desired uniformity of the attribute distributions was not achieved by the text-to-image model in the absence of DebiasPI's choice counting. EMD and JS-Div analysis shows that attempting to balance gender and race (right) results in greater bias compared to focusing solely on race (left) (blue: EMD 0.04, JS-Div 0.02 (left) and EMD 0.06, JS-Div 0.06 (right)).

so if it draws race categories uniformly, the probability that a specific race will be chosen in each trial is 0.56 i.e., $p = (\binom{9}{5} - \binom{8}{5})/\binom{9}{5}$, meaning in expectation there will be $n \times p = 40 \times 0.56 \approx 22$ images per race in the 200 generated images (or, $11 \pm \sigma\%$ for each race where $\sigma = \sqrt{n \times p \times (1-p)} \approx 3$). A simulation was run to confirm this, which showed that if the model attempted to draw race categories uniformly, then we should see between 8-14% images for each category. When comparing our experimental results in Fig. 7 (blue vs. orange distributions), we can see that DALL-E 3 indeed made an attempt to decrease over-representation of certain races while increasing the representation of under-represented ones. For example, Fig. 7(right) shows a notable decrease of White and an increase of Hispanic, Middle Eastern or North African, South Asian, and East Asian. However, Southeast Asian, Native Hawaiian and Other Pacific Islander and Indigenous populations remain under-represented. Analysis with JS-Div and EMD found that attempting to balance more than one set of attributes at the same time, here gender and race, results in greater bias compared to focusing solely on one type of attribute (race or gender).

Age. With the two ViT models (see Sec. 3.3), we obtained age estimates that placed the most common age of all generated images in the age group [34-64] and very few individuals in age group [65+]. One model [19], however, estimated that the remaining individuals ($\approx 40\%$) were in the age group [19-34], while the other model [14] estimated them to be in the age group [<19]. The majority of individuals in images inspected manually were young adults [19-34].

Skin Color. With the next set of experiments, we studied how we can instruct the AI model to generate images with a wide range of skin tones, now allowing DebiasPI to count attribute choices. Results for iterating on DebiasPI 50 times

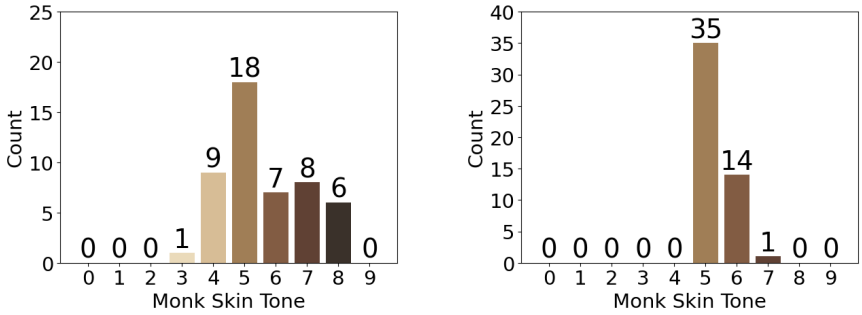

Fig. 8: Skin tone distributions over 10-choice Monk scale for attribute list 5-skin-tones (left) and attribute list 9-race choices (right).

with the 5-choice skin-tone list [dark tan, pale tan, purely black, warm brown, white fair] and the desired distribution uniform are shown in Fig. 8 (left). If the attribute list is race and not skin tone, DebiasPI generates skin tone that is only medium brown (Fig. 8 right). If DebiasPI has an input attribute distribution that is severely skewed toward light skin color, we found that DALL-E 3 was still not able to generate the palest four shades of the Monk scale. We acknowledge that these results were obtained in a relatively small experiment involving only 50 generated images but suggest that similar outcomes would be measured in larger experiments.

Desired Non-uniform Distributions. Although our focus has been on desired distributions that are uniform, we must note that DebiasPI can handle any desired input distribution. For example, when we asked DebiasPI to generate 90% female and 10% male characters in a 50-image experiment, it produced the desired 45 female and 5 male character images, as verified by an external classifier of gender. Another example involves an experiment with different desired race distributions (Table 1). Here, we first provided a 9-race attribute list without any distribution (Table 1) row 1). The text-to-image model most frequently selected the category "Black," a result aligned with the first cycles of DebiasPI of the experiment shown in Fig. 6. We then asked DebiasPI to produce a desired distribution that is uniform, which yielded in 5 or 6 images per race category (Table 1, row 2). Finally, we asked DebiasPI to produce 90% "White" faces and distributes the remaining 10% selections uniformly among the other 8 races. This resulted in 40 images in the "White" race category, and one or two images in the other categories (Table 1, row 3).

5 Conclusions

We proposed DebiasPI, an inference-time framework for robust attribute distribution control. With DebiasPI, a user can generate a series of images with attributes aligned to a target distribution, such as uniform distribution for fairness or a distribution that stresses specific traits. We envision, as a use case, a

Table 1: Race Distribution Obtained by DebiasPI for 50 images, without choice counting (row 1), with choice counting and a desired uniform distribution (row 2), and with choice counting and a desired non-uniform distribution (row 3). DebiasPI reaches the set distribution targets exactly (rows 2 and 3). Abbreviations: ME/NA (Middle Eastern or North African), NH/PI (Native Hawaiian or Pacific Islander), SE (Southeast).

Desired Distribution	Black	East Asian	Hispanic	Indigenous	ME/NA	South Asian	NH/PI	SE Asian	White
None	10	5	7	4	7	7	1	4	5
Uniform	6	5	5	6	6	6	5	5	6
Non-uniform	1	1	2	1	1	1	2	1	40

newsroom editor who might want to select among a diverse set of images of athletes. Our experiments show that DebiasPI is successful in generating images for representation of race and gender according to the desired attribute distribution.

Limitations of our work include the relatively small numbers of experimental data, the dependence of DebiasPI on the text-to-image model's ability to generate certain attributes (for example, skin tone), and the challenge that the model's internal beliefs or the external classifier's attribute analysis may not be entirely reliable, complicating DebiasPI's control over outputs.

The datasets we here publish may serve as benchmark comparisons by others in future work. Additional experiments with body-type, age, profession, and not-yet-explored attributes like affect would be interesting. Future work will also study ethical intervention when the text-to-image model is challenged with abstract concepts in the text to be visualized.

Acknowledgements

This work is supported in part by the U.S. NSF grant 1838193.

References

1. Browse by Career Cluster (2024), https://www.onetonline.org/find/career?c=0
2. Bansal, H., Yin, D., Monajatipoor, M., Chang, K.W.: How well can text-to-image generative models understand ethical natural language interventions? In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 1358–1370. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.88, https://aclanthology.org/2022.emnlp-main.88
3. Barrett, A.: Standards around generative AI. The Associated Press (Aug 2023), https://blog.ap.org/standards-around-generative-ai

4. Berelson, B.: Content Analysis in Communication Research. New York: Free Press (1952)
5. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., Ramesh, A.: Improving image generation with better captions, /https://cdn.openai.com/papers/dall-e-3.pdf
6. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 1493–1504 (Jun 2023). https://doi.org/https://doi.org/10.1145/3593013.3594095
7. Boeree, C.G.: Race. https://webspace.ship.edu/cgboer/race.html (2007), [Accessed 3-13-2024]
8. Casper, S., Davies, X., Shi, C., Gilbert, T.K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E.J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., Hadfield-Menell, D.: Open problems and fundamental limitations of reinforcement learning from human feedback. https://arxiv.org/abs/2307.15217 (2023)
9. Cho, J., Zala, A., Bansal, M.: DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers. CoRR **abs/2202.04053** (2022), https://arxiv.org/abs/2202.04053
10. Cho, J., Zala, A., Bansal, M.: DALL-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In: International Conference on Computer Vision (ICCV), Paris, France. pp. 3033–3054 (Oct 2023)
11. Clemmer, C., Ding, J., Feng, Y.: PreciseDebias: An automatic prompt engineering approach for generative AI to mitigate image demographic biases. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8581–8590 (2024), 10.1109/WACV57701.2024.00840
12. Ghosh, S., Caliskan, A.: 'Person' == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion". In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 6971–6985. Association for Computational Linguistics (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.465
13. Henry, F., Herwindiati, D.E., Mulyono, S., Hendryli, J.: Sugarcane land classification with satellite imagery using logistic regression model. IOP Conference Series: Materials Science and Engineering **185**, 012024 (6 pages) (Mar 2017). https://doi.org/10.1088/1757-899X/185/1/012024, https://dx.doi.org/10.1088/1757-899X/185/1/012024
14. Iakubovskyi, D.: Vit-age-classifier (dima806, version 12) (2024), https://huggingface.co/dima806/facial_age_image_detection
15. Kim, S., Cho, S., Kim, C., Lee, D., Baek, W.: minDALL-E on conceptual captions (2021), https://github.com/kakaobrain/minDALL-E
16. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023), https://arxiv.org/pdf/2304.08485
17. Monk, E.: Monk skin tone scale (2019), https://skintone.google
18. O'Connor, C., Joffe, H.: Intercoder reliability in qualitative research: Debates and practical guidelines. International Journal of Qualitative Methods **19**, 1609406919899220 (2020). https://doi.org/10.1177/1609406919899220, https://doi.org/10.1177/1609406919899220

19. Raw, N.: ViT-Age-Classifier (revision 461a4c4) (2023). https://doi.org/10.57967/hf/1259, https://huggingface.co/nateraw/vit-age-classifier
20. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA. pp. 10684–10695. CVPR, IEEE (Jun 2022)
21. Rozado, D.: The Political Biases of ChatGPT. Social Sciences **12**, 148 (03 2023). https://doi.org/10.3390/socsci12030148
22. Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S.L., au2, H.D.I., Dodge, J., Evans, E., Hooker, S., Jernite, Y., Luccioni, A.S., Lusoli, A., Mitchell, M., Newman, J., Png, M.T., Strait, A., Vassilev, A.: Evaluating the social impact of generative AI systems in systems and society. https://arxiv.org/abs/2306.05949 (2023)
23. Srinivasan, R., Uchino, K.: Biases in Generative Art - A Causal Look from the Lens of Art History. CoRR **abs/2010.13266** (2020), https://arxiv.org/abs/2010.13266
24. Stanley-Becker, I., Nix, N.: Fake images of Trump arrest show 'giant step' for AI's disruptive power (Mar 2023), https://www.washingtonpost.com/politics/2023/03/22/trump-arrest-deepfakes/
25. Sun, L., Wei, M., Sun, Y., Suh, Y.J., Shen, L., Yang, S.: Smiling women pitching down: Auditing representational and presentational gender biases in image-generative AI. Journal of Computer-Mediated Communication **29**(1) (Feb 2024). https://doi.org/10.1093/jcmc/zmad045, https://doi.org/10.1093/jcmc/zmad045
26. Wang, P.: DALLE-pytorch (2021), https://github.com/lucidrains/DALLE-pytorch
27. Young, J.: U.S. News Launches Generative AI Search Across USNews.com. U.S. News (Nov 2023), https://www.usnews.com/info/blogs/press-room/articles/2023-11-21/u-s-news-launches-generative-ai-search-across-usnews-com
28. Zack, T., Lehman, E., Suzgun, M., Rodriguez, J.A., Celi, P.L.A., Gichoya, P.J., Jurafsky, P.D., Szolovits, P.P., Bates, P.D.W., Abdulnour, P.R.E.E., Butte, P.A.J., Alsentzer, E.: Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study (2024), https://www.thelancet.com/journals/landig/article/PIIS2589-7500(23)00225-X/fulltext
29. Zhao, J., Khashabi, D., Khot, T., Sabharwal, A., Chang, K.W.: Ethical-advice taker: Do language models understand natural language interventions? In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 4158–4164. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.364, https://aclanthology.org/2021.findings-acl.364
30. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. arXiv preprint arXiv:2112.03109 (2021), https://github.com/FacePerceiver/FaRL