PreciseDebias: An Automatic Prompt Engineering Approach for Generative AI to Mitigate Image Demographic Biases

Colton Clemmer, Junhua Ding, Yunhe Feng
University of North Texas, Denton, TX, USA
coltonclemmer@my.unt.edu, junhua.ding@unt.edu, yunhe.feng@unt.edu

Abstract

Recent years have witnessed growing concerns over demographic biases in image-centric applications, including image search engines and generative systems. While the advent of generative AI offers a pathway to mitigate these biases by producing underrepresented images, existing solutions still fail to precisely generate images that reflect specified demographic distributions. In this paper, we propose PreciseDebias, a comprehensive end-to-end framework that can rectify demographic bias in image generation. By leveraging fine-tuned Large Language Models (LLMs) coupled with text-to-image generative models, PreciseDebias transforms generic text prompts to produce images in line with specified demographic distributions. The core component of PreciseDebias is our novel instruction-following LLM, meticulously designed with an emphasis on model bias assessment and balanced model training. Extensive experiments demonstrate the effectiveness of PreciseDebias in rectifying biases pertaining to both ethnicity and gender in images. Furthermore, when compared with two baselines, PreciseDebias illustrates its robustness and capability to capture demographic intricacies. The generalization of PreciseDebias is further illuminated by the diverse images it produces across multiple professions and demographic attributes. To ensure reproducibility, we will make PreciseDebias openly accessible to the broader research community by releasing all models and code.

1. Introduction

Images, a ubiquitous medium capturing visual narratives, profoundly shape human perception, beliefs, and decision-making processes. Ensuring the demographic impartiality of these images is crucial for both stakeholders and image-driven applications. Take, for instance, the potential perpetuation of stereotypes in image search results, which can inadvertently reinforce societal biases. Imagine a young girl searching for "CEO" on an image platform. If the top results predominantly showcase male CEOs, she might be mistakenly led to believe that the CEO role is exclusive to men. Such a lack of diverse representation not only threatens the fairness of image search outcomes but also casts doubt on their accuracy [5].

Beyond image search engines, machine learning models have consistently encountered challenges with demographic biases in images. Typically, these biases stem from the nature of their training datasets and the training methodologies utilized. A recent study by Cornell Data Science Team [25] highlighted these challenges in the context of the text-to-image model, DALL-E [19]. When the prompt "A photo of a doctor" was input into DALL-E, it was found to reinforce gender representation biases. Specifically, a male-to-female ratio of 2.35 was produced, diverging from the real-world ratio of 1.78. Further exploration with other gender-neutral prompts unveiled similar biases across various scenarios in text-to-image generation models [3, 14].

To resolve the statistical image demographic biases in image search and generation models, many approaches have been proposed. One prevalent approach involves the use of re-ranking algorithms, such as those detailed in [5, 31]. These algorithms enhance representation by prioritizing images of underrepresented groups in search results. However, while they can rearrange, they cannot inherently increase the diversity of the images, as they do not introduce new underrepresented images. Cornell Data Science Team [25] took a different approach by tweaking the input prompts with explicit gender specifications. Consequently, the model's output mirrored the specified gender. This method aligns with prior studies on prompt engineering for text-to-image models, suggesting that the integration of certain "modifiers" can steer the characteristics of the generated image [15].

Specifically, prompt engineering has been widely adopted to refine demographic nuances in generic prompts, facilitating the generation of images sensitive to demographic distinctions. As depicted in Figure 1, a generic prompt such as "A *nurse* wearing a blue smock is offering words of encouragement to a patient in an ICU" tends to generate images featuring a nurse of White ethnicity within an ICU setting. However, when the term *nurse* is refined

(a) *Generic* prompt (b) *Hispanic* prompt (c) *Black* prompt (d) *Asian* prompt (e) *White* prompt

Figure 1. Images are generated using varied prompts - "A *nurse* wearing a blue smock is offering words of encouragement to a patient in an ICU," where the term *nurse* can be further specified by ethnicity, including *Hispanic nurse*, *Black nurse*, *Asian nurse*, or *White nurse*, to guide the representation of ethnicity in the generated image.

with specific ethnic descriptors like *Hispanic, Black, Asian*, or *White*, the resulting images align with the intended ethnic representation. This observation underscores the pivotal role of refining prompts to mitigate demographic biases in image generation.

There exist many methods to enhance demographic characteristics in prompts, such as by substituting *nurse* with specific ethnic terms. A straightforward technique is using rule-based regular expressions to identify and replace a given subject (like *nurse*) with augmented variations (e.g., *Hispanic nurse*). However, this approach is limited to matching a predefined set of subjects. Furthermore, due to the intricate nature and nuances of ethnic descriptors, the rule-based system may erroneously detect ethnic subjects in the original prompts. For instance, integrating terms like *white* or *black* for ethnicity could lead to misconceptions about the presence of ethnic details in a prompt, considering these terms have different implications. One might suggest using Named Entity Recognition (NER) to pinpoint potential subjects of interest in a sentence, leveraging Python libraries such as SpaCy [8]. However, even the most proficient NER models have limitations, with top-performing models reaching only about ninety percent accuracy [16].

For these reasons, it is worth leveraging a pre-trained large language model (LLM) to fine-tune original prompts to augment their demographic specificity. The LLM possesses the ability to understand the context of word usage and assess whether a given sentence is statistically probable, given its comprehensive pre-training and advanced attention mechanisms as outlined by [28]. Consequently, LLMs can produce more precise and contextually natural prompts with demographic nuances compared to approaches reliant on rules or NER.

In this paper, we carefully fine-tune open-source LLMs to transform generic prompts devoid of demographic specifics into those that are demographically informed. More importantly, we propose a mechanism, titled PreciseDebias, that autonomously refines the prompt without the need for human intervention, ensuring it aligns seamlessly with the desired demographic feature distributions to match the intended target distributions accurately. The PreciseDebias framework aims to address the inherent statistical bias in a frozen machine learning model with a textual interface, harnessing the linguistic comprehension of both the frozen model and the LLM. By training the LLM, we enable it to adapt the original prompt to reflect particular demographic groups at expected probability rates, guiding the inherently biased image generation model towards more statistically representative demographic outputs. Figure 2 showcases an example of how fine-tuned LLMs work in conjunction with text-to-image models. Upon specifying a general prompt and the desired demographic ratio, these refined LLMs are capable of crafting demographic-specific prompts that follow the expected distributions. Subsequently, these enhanced prompts can be employed to produce the desired images.

We summarize the contributions of this paper as follows:

- We propose PreciseDebias, a novel end-to-end framework designed to transform generic prompts into demographically-tailored ones, ensuring the creation of images that follow desired demographic distributions.
- Our approach consists of novel algorithms that seamlessly integrate prompt engineering, model bias assessment, and balanced model training to enhance the capabilities of demographic-wise LLMs.
- Extensive experiments demonstrate the effectiveness, robustness, and generalization of the proposed PreciseDebias in rectifying demographic biases in prompt and image generation.
- To ensure reproducibility, we will release all associated models and training methodology codes upon the acceptance of this paper.

2. Related Work

This section briefly reviews the existing works in general bias mitigation, text-to-image model bias, and prompt engineering specific to Large Language Models (LLMs).

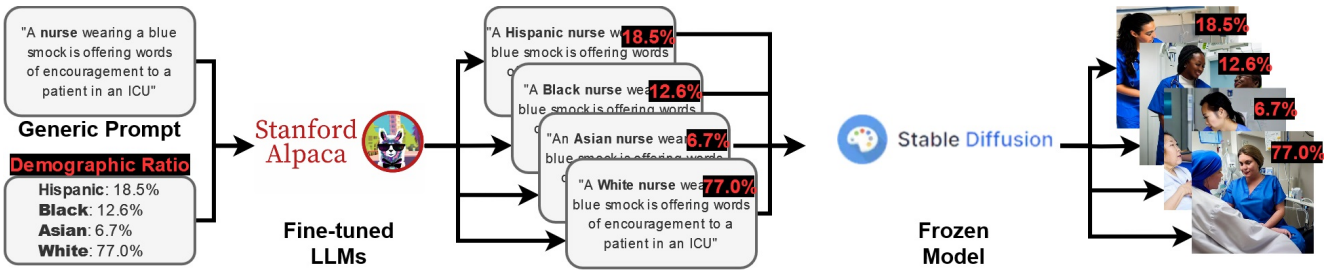


Figure 2. Inference workflow of the PreciseDebias framework. Given a generic prompt and specific demographic distributions, the PreciseDebias framework generates corresponding prompts and images that adhere to the desired ratios. A key component of this process is the fine-tuning of the involved Large Language Models (LLMs). Detailed insights into this mechanism can be found in the Section Methodology.

General Bias Mitigation. Bias mitigation has been an ongoing effort for decades now. The companies with the most funding and revenue have the highest impact on society and popular culture, yet even they struggle to overcome bias in their output [5]. Bias mitigation techniques, based on the stage of application, can be broadly categorized into: pre-processing, in-processing, and post-processing. Pre-processing methods, such as [10, 12, 17, 18, 30, 33], focused on diversifying and augmenting the training data to reduce biases. In-processing methods, such as [2,21,32] introduced fairness-aware penalties into the model training process to mitigate inherent biases. Post-processing methods adjusted and refined the model's post-training predictions, aiming to reduce biases. Relevant studies, including [5, 11, 13], modified or fine-tuned predictions based on specific fairness criteria after the model has been trained. The method proposed in this paper is classified within the post-processing category. Unlike traditional approaches which may diminish model performance due to repetitive input retention [22], our strategy generates fresh, high-quality data directly, thereby circumventing this issue.

Text-to-Image Model Bias. Text-to-Image models, such as DALL-E [19] and diffusion models [23] have had exponential growth over the last few years [20]. However, implicit biases present within their training datasets have been observed to influence model outputs. The study [3] demonstrated that inherent biases in text-to-image models, such as Stable Diffusion [20], can sometimes surpass the biases present in their training data. Bias in text-to-image models has been bench-marked by determining the diversity of output when given an ambiguous prompt. For example, a recent study [1] discovered that by embedding specific clues within a prompt to indicate an expectation of diversity, the model was more likely to yield a varied range of results. They also showed an effective method of evaluating the bias of one particular gender or ethnicity over another which is similar to our method. However, they regarded bias as any deviation from representing all ethnicities equally which is not a practically feasible goal when training given the wide range of human ethnicities worldwide. In contrast, our method focuses on specific ethnicities pertinent to a particular and practical use case, training the model to achieve diversity based on the specific needs and statistics desired by the developer.

Instruction-Following LLMs and Prompt Expansion. Recent studies have demonstrated the potential of large language models (LLMs) in proficiently following instructions [29]. Stanford's Alpaca project [24] exemplifies this capability through its LLaMA model [26]. This fine-tuned model can follow straightforward instructions, a feature we leverage by prompting the model to augment the original text. However, it is noteworthy that simply prompting the model to include ethnicities in responses can inadvertently lead to the over-representation of certain groups and a continuation of statistical bias in the opposite direction [6]. Prior research has also explored expanding text-to-image prompts [7] using LLMs. They used a reinforcement learning algorithm to maximize aesthetic appeal while maintaining prompt cohesion. Although this methodology provided foundational insights for our initial fine-tuning and evaluation approach, their specific optimization technique and loss function did not align with our objectives. In addition, their fine training method, consisting of several machine learning models working in tandem, was too resource intensive for our purposes.

3. Methodology

In this section, we begin by providing an overview of PreciseDebias, followed by a detailed discussion of its integral components.

3.1. PreciseDebias Overview

PreciseDebias is mainly composed of two core components: fine-tuned Large Language Models (LLMs) and frozen text-to-image models, as illustrated in Figure 2. The fine-tuned LLMs have the capability to transform generic prompts, combined with specified demographic distribu-

tions, into augmented prompts that accurately mirror the anticipated demographic ratios. The term "frozen text-to-image model" indicates its adaptability. Essentially, it can be any model that takes a text input and generates corresponding images. In the following subsections, we delve into the intricate designs of both these components.

3.2. Fine-tuning Large Language Models

We first briefly introduce the notations of involved subjects, e.g., generic prompts, demographic characteristics, and LLMs. Then we present the details about bias measurement, model training, and model validation.

3.2.1 Preliminaries

Let i denote a user-provided input to the LLM (i.e., a generic prompt), and let c represent the demographic characteristic intended to be augmented by the fine-tuned LLM, denoted as M. We express the set of all generic input prompts and the set of all demographic characteristics as I and C, respectively. Given any $i \in I$, our objective is to integrate each $c \in C$ into the final demographic-specific output of LLM M. Consequently, we aim to produce an output prompt set O, which encompasses all possible combinations of elements drawn from sets I and C.

$$O = \{O_c^i \mid i \in I, c \in C\} \tag{1}$$

where O_c^i represents the demographic-aware output prompt created by enhancing the generic prompt i with the demographic characteristic c.

For each demographic characteristic $c \in C$, there is a corresponding expected ratio associated with any output of LLM M that incorporates c. We refer to these expected ratios as the ground truth and represent them by the set $G = \{G_c \mid c \in C\}$, where G_c denotes the expected ratio for the demographic characteristic c. It is important to note that, in certain contexts, the subject of a sentence may identify with multiple demographic characteristics, such as in the case of ethnicity. Consequently, the sum of the percentages for each characteristic may exceed 100%[1].

LLMs typically require pairs of prompts and their associated outputs to construct a training dataset for model fine-tuning. Let the dataset D represent the collection of all tuples consisting of a generic prompt $i \in I$ and its corresponding expected output $O_c^i \in O$. Consequently, the dataset D can be formulated as follows:

$$D = \{(i, O_c^i) \mid i \in I, c \in C\} \tag{2}$$

[1]NOTE: Estimates for the above race groups (White, Black or African American, and Asian) do not sum to totals because data are not presented for all races. Persons whose ethnicity is identified as Hispanic or Latino may be of any race. - U.S. Bureau of Labor Statistics

Let us consider a function $\mathcal{F}$ that fine-tunes a pre-trained LLM M based on the dataset D and a set of parameter configurations S. This results in an improved version of the model, denoted as M', capable of generating augmented prompts that adhere to the ground truth demographic distributions G for each demographic characteristic $c \in C$.

$$\mathcal{F}(M, D, G, C, S) \rightarrow M' \tag{3}$$

To further clarify the above notations, we explain them using the example depicted in Figure 2. We present just one generic input, namely i, which is: "A nurse wearing a blue smock is offering words of encouragement to a patient in an ICU." However, PreciseDebias can simultaneously tune a set of generic prompts. The demographic characteristics featured in Figure 2 can be expressed as $C = \{$*Black, Hispanic, Asian, White*$\}$. When c is specified as *Black*, the resulting prompt O_c^i becomes: "A Black nurse wearing a blue smock is offering words of encouragement to a patient in an ICU," and the corresponding expected ratio G_c is set at 12.6%.

3.2.2 Training Data Generation

With the popularity of GPT-3 API [4], synthetic data has become an increasingly common method of training models [24]. In this paper, we employ instruction-following LLMs to produce the training dataset D, as defined in Equation 2. First, we adopt a few-shot approach [4] to prompt GPT-3 to create additional prompts that incorporate specific demographic characteristics. We will illustrate this process using ethnicity as an example of how to craft instructions to guide the LLMs. Specifically, by providing GPT-3 with a few hand-crafted prompts (typically five to ten), we design the following instructions to generate more prompts that integrate ethnicity information for a text-to-image model.

> Write a few prompts for a text-to-image model that describes a person in a specific occupation. Try to include these ideas in the prompts: diversity, specificity, imagination, detail, and emotion. Use language that paints a picture and describes a story. Always include the ethnicity of the person in the prompt.

The next step is to obtain the generic prompts by obscuring the ethnicity information in the prompts generated by GPT-3. Specifically, we compile a list of all words related to ethnicity from our dataset and used regular expressions to remove them. Finally, to create generic and demographic-specific prompt pairs, we utilize the Alpaca instruction template [24] to rewrite and enrich the generic prompt with ethnicity information. Below, we provide an example of the Instruction, Input, and Response formats used in Alpaca.

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.
Instruction: Expand this text-to-image prompt to include a specific ethnicity for the subject.
Input: A female nurse is caring for a premature baby in an incubator in the neonatal unit.
Response: A female Hispanic nurse is caring for a premature baby in an incubator in the neonatal unit.

3.2.3 Model Bias Measurement

As the aim of PreciseDebias is to mitigate biases in a precise manner, it is critical to measure the bias present in the outputs generated by LLM M. To this end, we utilize the notation $E = \{E_i \mid i \in I\}$ to represent the set of outputs generated by an LLM M in response to a generic prompt set I. Here, each E_i corresponds to a collection of n outputs generated for a particular prompt $i \in I$.

$$E_i = \{E_i^j \mid j \in [1, n], i \in I\} \tag{4}$$

where E_i^j represents the j-th generation corresponding to input i. Each generated output E_i^j is assumed to contain a member of the demographic characteristic set C. Let A_c denote the total number of appearances of demographic characteristic $c \in C$ across all outputs in E.

$$A_c = \sum_{i \in I} \sum_{j=1}^{n} \begin{cases} 1 & \text{if } E_i^j \text{ contains } c \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Once the model has been fine-tuned, its bias can be assessed in comparison to the ground truth set G. Therefore, the probability of the demographic characteristic c appearing in response to a given input i can be calculated as the ratio of A_c to the total number of generations, $|E|$.

$$P(c|M) = \frac{A_c}{|E|} \tag{6}$$

Then we can calculate the statistical bias Δ_c for a given demographic characteristic c by taking the difference between the ground truth probability G_c and the measured model probability $P(c|M)$.

$$\Delta = \{G_c - P(c|M) \mid c \in C\} \tag{7}$$

where O_c^i represents the demographic-aware output prompt created by enhancing the generic prompt i with the demographic characteristic c.

The entire procedure of measuring the bias Δ for the generated outputs of LLM M is illustrated in Algorithm 1. Note that we discard all negative Δ_c values because, given our training method, we may only train M toward a goal and not away from it. Furthermore, we can logically infer that if the probability of a certain outcome increases, the probabilities of other outcomes must also decrease.

Algorithm 1 Bias_Measure(M, I, G, C, n)

```
E ← ∅                                   ▷ Initialize E, the output of model M
for i in I do                           ▷ Generate n outputs for each prompt i
    for j ∈ [1, n] do
        E_i^j ← Generate output from M using i as input
    end for
end for
A ← ∅                                   ▷ Initialize A, the appearing count of c in E
for each c ∈ C do                       ▷ Calculate the appear. count of c
    Calculate A_c using Equation 5
end for
Δ ← ∅                                   ▷ Initialize Δ for all c ∈ C
for c in C do
    bias ← G_c − A_c/|E|                ▷ Calculate bias using Equation 7
    if bias > 0 then                    ▷ Only keep the positive bias
        Δ_c ← bias
    end if
end for
return Δ
```

3.2.4 Proportional Model Training

Based on the output bias Δ of the LLM model M, we can further fine-tune M by employing proportional learning rates tailored to modulate the learning speed across different demographic characteristics. Let R represent the set of learning rates, with each individual rate R_c corresponding to a positive adjustment factor Δ_c for every member of the demographic set C. To calculate each R_c, we start with a constant learning rate r, multiply it by the associated bias Δ_c for a specific demographic characteristic c, and then divide the result by the total number of items in the generic prompt set I.

$$R = \{\frac{r * \Delta_c}{|I|} \mid \Delta_c > 0\} \tag{8}$$

where $|I|$ is the size of the generic prompt set I and r is the constant learning rate.

To refine the model M, we generate a new dataset for each element within the set R. Specifically, each newly formed dataset D_c is a subset of the original dataset D, but is focused exclusively on the demographic characteristic c.

$$D_c = \{(i, O_c^i) \mid i \in I\} \subset D \tag{9}$$

Each new dataset is rerun through the fine-tuning process to adjust the probability of any output containing some demographic characteristic. This function of fine-tuning is denoted as $\mathcal{F}$.

$$\mathcal{F}(M, D, G, C, R) = \prod_{c \in C \& \Delta_c > 0} \mathcal{F}(M, D_c, G_c, \{c\}, R_c) \tag{10}$$

where $\prod$ represents iteration over every member of set C to produce an updated version of M. Algorithm 2 shows the detailed fine-tuning procedures.

Algorithm 2 Fine_Tune(M, I, D, Δ, C, r)

for c in C **do**
  if $\Delta_c > 0$ **then**
    $R_c \leftarrow r * \Delta_c / |I|$ ▷ $|I|$ is the size of input set I
    $M \leftarrow \mathcal{F}(M, D_c, G_c, \{c\}, R_c)$ ▷ Fine-tune M
  end if
end for
return M

After fine-tuning the model for all demographic characteristics that exhibited positive bias values, we proceed to re-evaluate model M using Algorithm 1. If every element of the resulting bias Δ falls below a pre-established threshold t, we will terminate the process. If not, we will reapply Δ to $\mathcal{F}$ and continue to assess the model's bias iteratively until it is reduced below the threshold. Finally, we will obtain a model M, which is capable of generating augmented prompts that adhere to the ground truth demographic distributions G for each demographic characteristic $c \in C$. The entire procedure is illustrated in Algorithm 3.

Algorithm 3 Precise_Debias(M, I, G, C, n, r, t)

loop
  $\Delta \leftarrow Bias\_Measure(M, I, G, C, n)$ ▷ Alg. 1
  $pass \leftarrow True$
  for Δ_c in Δ **do**
    if $\Delta_c > t$ **then**
      $pass \leftarrow False$ ▷ Ensure all Δ_c values are < t
    end if
  end for
  if $pass = True$ **then** ▷ M is done
    return M
  end if
  $M \leftarrow Fine\_Tune(M, I, D, \Delta, C, r)$ ▷ Alg. 2
end loop

3.3. Frozen Text-to-Image Models

The output of fine-tuned LLM M will serve as the input prompt for the text-to-image models to generate images. As a downstream component within PreciseDebias framework, the text-to-image model is designed to be a plug-and-play component. Consequently, we can employ various text-to-image models, such as Stable Diffusion [20] or DALL-E [19], to convert text into images. In this paper, we have chosen to utilize Stable Diffusion for the task of transforming text into images. However, as we have discussed above, PreciseDebias is highly adaptable, capable of accommodating a wide range of alternatives that offer a text-based interface for image generation.

4. Evaluation and Experimental Results

In this section, we present the experimental settings and results to validate the effectiveness of PreciseDebias .

4.1. Experiment Settings

We select LLaMA-7B [26] as the pre-trained LLM model to create the unbiased prompts, and the Stable Diffusion [20] to generate images according to the unbiased prompts. But PreciseDebias is also flexible enough to incorporate other LLMs like LLaMA 2 [27] and text-to-image models like Midjourney[2].

To fine-tune LLaMA-7B more efficiently, we utilize the Low-Rank Adapter (LoRa) [9] to reduce the required training memory on four V100 32GB GPUs. Specifically, we train the LoRa adapter over the dataset containing 560 prompts generated by the approaches proposed in Subsection 3.2.2. Detailed source code is available on GitHub. [3].

4.2. Model Validation

In order to validate our proposed PreciseDebias method, we utilize Stable Diffusion [20] for local text-to-image generation. For every generic prompt given, we generate 45 images utilizing Stable Diffusion, yielding five images per run over nine runs. Considering that demographic characteristics like ethnicity can be challenging to identify [5], we manually detect demographic attributes present in the synthesized images. To evaluate the likelihood of each demographic characteristic's occurrence, we aggregate the instances of each demographic feature from all generated images and then divide by the total number of human-centric images. On completion of our validation dataset analysis, we obtain an assessment of the statistical biases present in the combined system of the trained language model and the frozen text-to-image model.

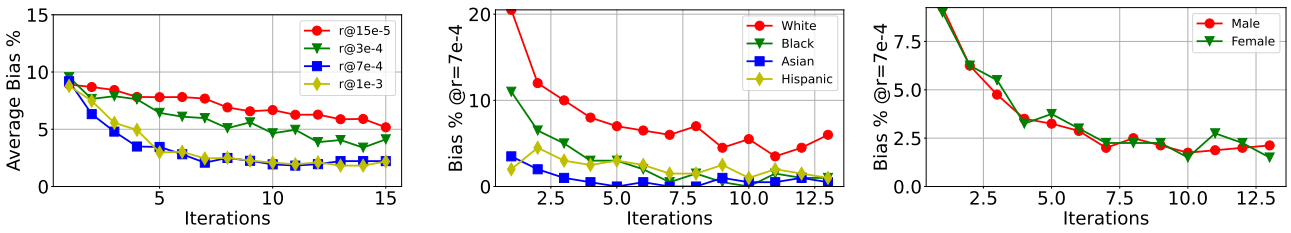
4.3. Experimental Results

We first summarize the debiasing performance of PreciseDebias, and then compare it with several baselines.

4.3.1 Debiasing Performance

The hyperparameter constant learning rate r plays a pivotal role in PreciseDebias . We initiate our study by examining the influence of r on the debiasing performance. A large learning rate usually results in faster training but may also increase the risk of convergence to a local maximum for a particular demographic characteristic. On the other hand, a

[2] https://www.midjourney.com/
[3] github.com/ResponsibleAILab/Precise_Debias



(a) Debiasing with different learning rates (b) Ethnicity debiasing (c) Gender debiasing

Figure 3. (a) Average debiasing performance across all ethnicities and genders for different learning rates. (b) Debiasing performance of each ethnicity over every iteration. (c) Debiasing performance of each gender over every iteration. All y-axis values in the above figures represent the percentage difference between the demographic ratio of generated images and the ground truth.

	White	Black	Asian	Hispanic	Overall
# of images	2974	595	280	638	4487
# of successes	2920	590	280	590	4380
# of fails	54	5	0	48	108
Success rate	98.2%	99.2%	100%	91.9%	97.6%

Table 1. Success rate of image generation corresponding to augmented ethnicity prompts.

	Male	Female	Overall
Number of Images	1404	1462	2866
Number of successes	1400	1460	2860
Number of failures	4	2	6
Success rate	99.8%	99.9%	99.9%

Table 2. Success rate of image generation corresponding to augmented gender prompts.

learning rate that is too low will result in unnecessary computational overhead and longer training time than necessary. Figure 3a illustrates how varying values of the learning rate contribute to bias mitigation. Optimal and consistent debiasing is observed when r is set to $7e{-}4$. When increasing r to $1e{-}3$, it results in a less competitive debiasing outcome during the initial five epochs. In the following experiments, we maintain the learning rate at $7e{-}4$ unless otherwise specified.

In our study, we evaluate the debiasing performance of PreciseDebias concerning two demographic characteristics: ethnicity and gender. Figure 3b illustrates the breakdown reduction in the percentage difference between the ethnicity ratios of the generated images and the ground truth as training iterations increase. Similarly, Figure 3c shows how the discrepancy in gender distributions between the generated images and the ground truth narrows with an increased number of training interactions. In our experiments, we set the number of training iterations as 10 because it achieves the best performance for most demographic characteristics.

4.3.2 Demographic-specific Prompt Quality

To evaluate the quality of augmented demographic-specific prompts, we conduct a detailed analysis by determining the success rates at which the generated images reflect the targeted demographic characteristics. Table 1 reveals an overall success rate of 97.6% for generating ethnicity-specific images using the augmented prompts from PreciseDebias. With the exception of *Hispanic*, the success rates for all other ethnicities are above 98%. For characteristics with fewer distinct features, such as gender, Table 2 indicates success rates exceeding 99%.

4.3.3 Baseline Comparisons

We compare the performance of PreciseDebias with both regular expression based and Named Entity Recognition (NER) based solutions.

Regular expression approaches often lack the adaptability to interpret ambiguous demographic terms effectively. For instance, when confronted with generic prompts that use color terms such as *white* and *black*, these methods often misinterpret them as racial descriptors rather than simple color attributes. Consequently, they fail to bring about any change to the prompt. Consider the illustrative example in Figure 4. The prompt reads, "*A construction worker stands next to a white truck*". Here, *white* pertains to the truck's color and not racial identity. As illustrated in Figure 4a, regular expression methods leave the original prompt unaltered, mistakenly identifying racial implications. This results in all generated images portraying a *white* truck but with all *white* construction workers. On the other hand, PreciseDebias can diversify the racial profiles in generated images by adeptly augmenting such generic text prompts (see Figure 4b for more details).

The rule-based prompt augmentation enabled by NER can also yield unforeseen outcomes. Specifically, NER methods can identify and extract entities related to individuals, such as names and roles, from raw text. Subsequently,

(a) Regular expression: all generated images are *White*

(b) PreciseDebias: *Hispanic, Black, Asian*, and *White*

Figure 4. Comparison of the debiasing performance between regular expression solutions and PreciseDebias regarding "*A construction worker stands next to a white truck.*"

rule-based approaches may incorrectly assign an ethnicity to the identified individual, resulting in nonsensical image generation. As shown in Figure 5, consider the generic prompt "*Jim from accounting sipped coffee next to the copy machine*". In this case, NER detects *Jim* as a person-related entity and generates the unusual prompt "*Hispanic Jim from accounting sipped coffee next to the copy machine.*" Furthermore, it is worth noting that even state-of-the-art NER models have limitations. Even top-performing models achieve only around ninety percent accuracy [16], potentially impacting prompt enhancement performance.

Figure 5. Named entity recognition (NER) falsely converts "*Jim from accounting sipped coffee next to the copy machine.*" to "*Hispanic Jim from accounting sipped coffee...*"

4.3.4 Generalization Across Diverse Professions

The proposed PreciseDebias showcases notable flexibility and generalization in enhancing demographic representations across various professions. It can generate images in line with designated demographic attributes in response to generic text prompts. In Table 3, we present the capability of PreciseDebias to diversify demographic portrayals across five distinct professions, namely chef, firefighter, doctor, teacher, and artist, spanning four ethnicities. As shown in the *Generic* row, absent the demographic-specific prompts, the default output tends to prioritize certain population groups, thereby exacerbating the disparity faced by underrepresented ethnic groups. The rows labeled *Hispanic, Black, Asian*, and *White* provide a glimpse into the images generated by PreciseDebias for the specified ethnicity. Note that PreciseDebias is also able to create a variety of demographically-wise images that align with the intended demographic distributions.

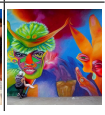
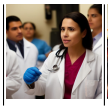
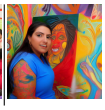
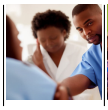
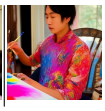
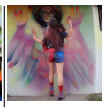
Ethnicity	Chef	Firefighter	Doctor	Teacher	Artist
Generic					
Hispanic					
Black					
Asian					
White					

Table 3. Generalization of PreciseDebias across various professions and ethnicities.

5. Conclusion and Discussion

This paper proposes PreciseDebias to produce demographic-specific images by converting generic prompts into demographically-informed ones, adhering to the desired distribution of demographic characteristics. To achieve this, we design an innovative prompt-tuning algorithm, empowering large language models (LLMs) with enhanced sensitivity to demographic nuances. These fine-tuned text prompts are then utilized by the frozen text-to-image model to generate images. Our comprehensive experiments validate the effectiveness of PreciseDebias in rectifying biases across various demographic attributes, including ethnicity and gender. Furthermore, the quality, robustness, and generalization of the proposed PreciseDebias and the resultant images have been demonstrated.

While we have categorized ethnicity and gender-related features into a predetermined set of categories in this paper, it is essential to acknowledge that demographic attributes in reality encompasses a broader and more intricate spectrum. For instance, gender should be more inclusive by not being confined to the binary distinctions of male and female.

Acknowledgment: The reported work was supported in part by the Microsoft Accelerate Foundation Models Research Grant.

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? *arXiv preprint arXiv:2210.15230*, 2022. 3

[2] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017. 3

[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 *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023. 1, 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[5] Yunhe Feng and Chirag Shah. Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11882–11890, 2022. 1, 3, 6

[6] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023. 3

[7] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*, 2022. 3

[8] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 2

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[10] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012. 3

[11] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019. 3

[12] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9572–9581, 2019. 3

[13] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *Ieee International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851, 2019. 3

[14] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. *arXiv preprint arXiv:2304.06034*, 2023. 1

[15] J Oppenlaender. A taxonomy of prompt modifiers for text-to-image generation (arxiv: 2204.13988). arxiv, 2022. 1

[16] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020. 2, 8

[17] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[18] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3, 6

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 6

[21] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *European Conference on Computer Vision(ECCV)*, pages 746–761, 2020. 3

[22] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arxiv:2305.17493*, 2023. 3

[23] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[24] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 3, 4

[25] Cornell Data Science Team, Aug 2022. 1

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3, 6

[27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[29] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 3

[30] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575, 2018. 3

[31] George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2532–2538, 2022. 1

[32] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, page 335–340, 2018. 3

[33] Yi Zhang and Jitao Sang. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In *Proceedings of the 28th ACM International Conference on Multimedia (MM)*, pages 4346–4354, 2020. 3