This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multi-Group Proportional Representation for Text-to-Image Models

Sangwon Jung[1] Alex Oesterling[2] Claudio Mayrink Verdun[2]
Sajani Vithana[2] Taesup Moon[1,3] Flavio P. Calmon[2]

[1] ECE, Seoul National University
[2] Harvard University [3] ASRI/INMC/IPAI/AIIS, Seoul National University

Abstract

Text-to-image (T2I) generative models can create vivid, realistic images from textual descriptions. As these models proliferate, they expose new concerns about their ability to represent diverse demographic groups, propagate stereotypes, and efface minority populations. Despite growing attention to the "safe" and "responsible" design of artificial intelligence (AI), there is no established methodology to systematically measure and control representational harms in image generation. This paper introduces a novel framework to measure the representation of intersectional groups in images generated by T2I models by applying the Multi-Group Proportional Representation (MPR) metric. MPR evaluates the worst-case deviation of representation statistics across given population groups in images produced by a generative model, allowing for flexible and context-specific measurements based on user requirements. We also develop an algorithm to optimize T2I models for this metric. Through experiments, we demonstrate that MPR can effectively measure representation statistics across multiple intersectional groups and, when used as a training objective, can guide models toward a more balanced generation across demographic groups while maintaining generation quality. [1]

1. Introduction

Recent text-to-image (T2I) generative models, such as Google Gemini [25], DALL-E [56], Midjourney [45], Adobe Firefly [8] and Stable Diffusion [58], have pushed the boundaries of creativity with AI, offering new possibilities in fields ranging from digital art to product design. However, as these technologies become more prevalent, there has been a surge in concerns about their portrayal of diverse population groups [11], their potential biases [16], and their ability to promote stereotypes [21, 42]. A notable example is the controversy surrounding Google's Gemini AI, which produced historically inaccurate images [57] and received significant attention in the press [55].

The challenge of achieving fair representation in generative AI requires, first and foremost, a precise methodology for *measuring* representation. However, current approaches that assess representation in T2I models face several limitations surrounding how they define and measure *representational fairness* [63]. Most studies [16, 23, 64] focus on a single axis of outward-facing identity representation, such as gender, overlooking the complex interplay of multiple identity factors. While some research attempts to consider multiple axes of representation, e.g., [11, 59, 65, 66], there is a lack of systematic measurement methods, particularly for intersectional groups. Determining appropriate representation goals is also challenging. Recent frameworks like BIGbench [40] and OpenBias [19] have proposed more systematic approaches to evaluate bias in text-to-image models by leveraging multi-modal LLMs. However, these methods still lack a theoretically grounded framework for measuring representational fairness across intersectional groups and optimizing generative models accordingly.

Our proposed framework addresses these limitations by providing a unified mathematical framework that can measure and mitigate representational harm. For instance, sometimes we may want equal representation (such as generating images of a CEO with equal gender representation), and in others, we may care about reflecting some real-world or historically accurate distribution of identities (such as not generating racially diverse 1943 German soldiers as done by Gemini [57]). *In this work, we apply a metric to T2I models that can address both of these needs and allow us to improve representation in image generation.* Our approach represents the comprehensive framework that can flexibly evaluate representation across various complexities of intersectional groups while being adaptable enough to accommodate both equal and proportional representation goals, providing a principled foundation for measuring and mitigating representational harm in T2I systems. Indeed, we empirically show that optimizing a generative image model

[1] The code is available at https://github.com/sangwon-jung94/mpr-t2i

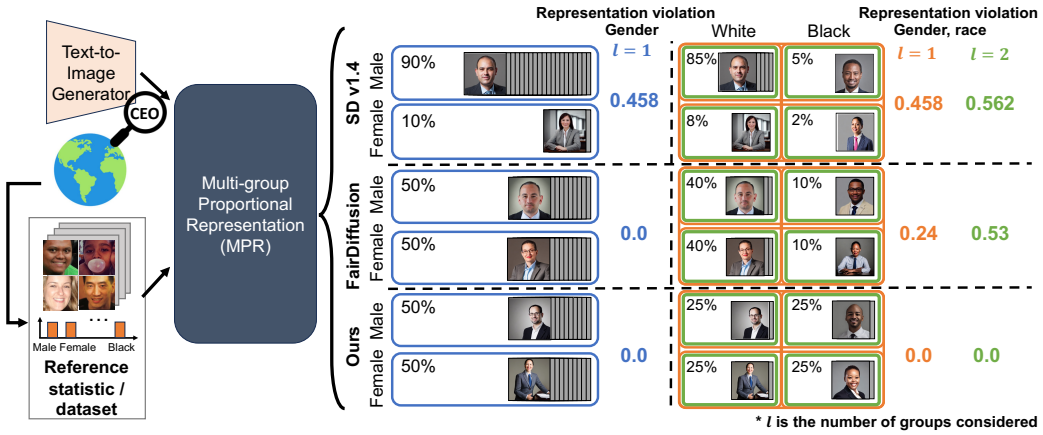


Figure 1. Overview of the Multi-Group Proportional Representation (MPR) framework for evaluating text-to-image generators. Given a prompt (e.g., "CEO"), the framework compares the distribution of demographic attributes (such as gender and race) between generated images and a reference dataset or target statistics. MPR measures the maximum deviation in representation across multiple intersectional groups, enabling systematic quantification of representational biases in generated images.

with MPR improves the representation of its generations (Fig. 1).

This paper contributes to the field of fairness in generative AI by introducing the use of Multi-group Proportional Representation (MPR), initially proposed for retrieval tasks [48], as a general metric for measuring and improving representation in T2I models. This extension provides a robust and flexible approach to quantifying fair proportional representation with respect to general computationally-identifiable groups. Importantly, our proposed MPR metric can be easily adjusted to accommodate complex or newly emerging definitions of group identity, making it a versatile tool for measuring representational harm in T2I models that can navigate the aforementioned tensions between factuality and representation. Our main contributions are as follows:

- We introduce the Multi-Group Proportional Representation (MPR) metric, providing the first theoretically grounded framework for measuring and optimizing for representational fairness in T2I models. While MPR was originally proposed for retrieval tasks, we show how it can effectively capture intersectional representation in generated content with remarkable flexibility in handling different demographic attributes, group definitions, and fairness criteria.
- We demonstrate the scalability and effectiveness of MPR through experiments with state-of-the-art T2I models, showcasing its ability to capture biases related to various facets of identity. In particular, we use MPR to identify critical representation gaps for popular models in professional representation, trait attribution, and disability portrayal. We additionally show that MPR scales to a large set of attributes and adapts to query-contextual representation evaluation.
- We demonstrate how MPR can be used as a training objective to optimize T2I models for fair representation. Our experiments show that MPR-based fine-tuning achieves a more balanced demographic representation than existing debiasing methods while maintaining image quality and semantic accuracy.

The paper is organized as follows. In Section 2, we present a comprehensive analysis of related work. In Section 3, we introduce MPR and discuss its theoretical properties in a generative context, including its ability to capture nuanced representational harm of intersectional groups. In Section 4, through extensive experiments, we demonstrate the practical effectiveness of MPR in capturing representational harm across various demographic groups and their intersectional identities. In Section 5, we illustrate how MPR can be used to fine-tune a text-to-image generation model to mitigate the harm.

2. Related Work

Representational harm in T2I systems. Recent research has highlighted key areas where T2I generation systems exhibit representational harm across various societal and demographic factors. One prominent area of concern is the representation of professions. Multiple studies have identified significant issues in how these systems depict various occupations, particularly related to gender [23, 31, 49], race [39], skin-tone [16, 66] or age [59, 66]. Some studies have shown that these issues extend to traits described by adjectives (*e.g.*, "poor") and certain nouns (*e.g.*, "thug") [11, 22]. Research has also identified unfair representation in disability [42] – including emotional expression and representation of objects like wheelchairs – and socioeconomic representation, where certain economic conditions are disproportionately

associated with specific groups [11, 21]. With an increasing understanding of the wide range of unfairness potentially present in T2I generation systems, [15] proposes a method to automate unfairness identification and dynamically generates potential areas of unfairness risk based on the prompt used to generate an image.

Fairness metrics and evaluation methods. Prior bias evaluation approaches in T2I models have various limitations, such as using qualitative evaluations with humans to assess the presence and impact of biases [9]. When examining a single type of bias, researchers often use a uniform distribution (e.g., each group represented equally) as a reference point for comparison [9, 16]. This approach provides a baseline against which deviations can be measured, helping to identify specific biases within the model outputs. However, the complexity of bias necessitates the consideration of multiple aspects simultaneously, as demonstrated in several studies [11, 39, 50, 59, 66]. Notably, quantitative metrics for assessing bias in intersectional groups remain limited, with only a few metrics currently available [20, 40, 59, 66]. However, these metrics lack a systematic framework for determining the level of granularity of intersectionality in measuring representation and for selecting appropriate representation targets.

Bias mitigation strategies. Recent research has explored various approaches to mitigate bias in T2I models. Ethical statement prompting [9] employs post-processing and manual prompt tuning by appending statements like "if all individuals can be a lawyer irrespective of their gender" to prompts. [23] introduced FairDiffusion, a post-processing method that modifies text guidance using a lookup table to add text guidance vectors of minor groups while subtracting those of major groups. DebiasVL [18], TIME [49] and UCE [24] utilize linear projections to filter out identity-related information from text embedding vectors. [37] developed FairMapping, another post-processing method using linear projection of text embedding vectors. [31] explored soft prompt tuning as an efficient training method for debiasing. More recently, some works proposed more sophisticated methods for debiasing by adjusting features fed into attention modules [51, 65], utilizing reinforcement learning [44] or optimal transport-based regularization [59]. [38] tilted the distribution of training and generated data through a pseudo-density to adjust the fidelity and diversity of generative models. While not specifically focused on T2I systems, [33] introduced DSM, an in-processing re-weighting learning scheme optimized for diffusion models, and [17] developed FairSwitch, a post-processing method that modifies samples in the middle of the backward process of diffusion models. Finally, [67] propose a method in predictive settings to scale fair learning to large numbers of intersectional groups using knowledge distillation.

3. Multi-group Proportional Representation in Image Generation

Our goal is to generate images that are accurate with respect to a given prompt while being representative across various intersectional population groups in a given social or historical context. As mentioned in the Introduction, we draw upon the MPR metric, first introduced in retrieval systems [48], and extend it to a T2I generation. The formal definition is provided in Definition 1.

Definition 1. Let $X_g \in \mathbb{R}^d$ for some $d < \infty$ be the random variable representing a generated image from a text-to-image generative model, with a distribution G_q, conditioned on a given prompt q. Let $X_r \in \mathbb{R}^d$ be the random variable representing an image from a reference population with a distribution R. For a bounded class of functions $\mathcal{C}$, where $c : \mathbb{R}^d \to \mathbb{R}$ for each $c \in \mathcal{C}$, the MPR metric is defined as:

$$\text{MPR}(\mathcal{C}, G_q, R) \triangleq \sup_{c \in \mathcal{C}} \left| \mathbb{E}_{G_q}[c(X_g)] - \mathbb{E}_R[c(X_r)] \right|. \quad (1)$$

For X_g and X_r in (1), one can use the direct pixel values, d-dimensional vector embeddings of the respective images, or attribute vectors indicating the sample's group membership with respect to the sensitive attributes being considered. In this work, we opt for the attribute vectors to conduct experiments with reduced noise in the data.

We note that MPR has appeared in different contexts under different names, e.g., it is also called the maximum mean discrepancy (MMD) [12, 26] of a selected set of representation statistics determined by $\mathcal{C}$, with respect to the generative and reference distributions. It can be seen as a special case of integral probability metrics (IPM) [47, 53, 61]. In fairness literature, MPR builds off of work in multicalibration [28] and multiaccuracy [32] for multi-group fair classification.

The MPR metric is a generalization of many of the representational fairness metrics for image generation used in prior studies. For instance, by selecting 1) $X_g, X_r \in \{0, 1\}$ as the binary membership in a single group type, such as gender or race, 2) uniformly distributed R over $\{0, 1\}$, and 3) $\mathcal{C} = \{c_0, c_1\}$, where $c_0(X) = \mathbf{1}_{\{X=0\}}$ and $c_1(X) = \mathbf{1}_{\{X=1\}}$ with $\mathbf{1}(\cdot)$ denoting the identity function, MPR computes the maximum difference between the probabilities across groups, as done in [23, 24]. Similarly, when X_g, X_r are vectors representing membership in multiple groups, and $\mathcal{C}$ corresponds to a set of full-depth decision trees, MPR measures unfairness across all possible intersectional groups as in [59, 66].

The key advantage of MPR over the existing metrics lies in its flexibility regarding the choice of reference distribution and function class $\mathcal{C}$. By employing context-aware reference distributions that adapt to the prompt, MPR can address the *diversity versus prompt relevance* tradeoff encountered previously by Gemini [57]. For example, if the prompt is *"an*

image of a 2022 FIFA World Cup athlete", a representation metric should not penalize the system for failing to generate images of female athletes, as this aligns with the real-world context of events that took place in 2022. To achieve this, we can measure MPR against a reference distribution R_q conditioned on the prompt, ensuring that the evaluation focuses only on a contextual and relevant reference distribution.

Additionally, MPR allows control over the granularity of representational difference measurements by adjusting the complexity of the function class. Some users may prefer to measure deviations between generated and reference distributions with high precision, using complex function classes like full-depth decision trees or deep neural networks. However, as the number of group attributes grows or when continuous embedding vectors are involved, it may become necessary to reduce the complexity. For instance, as in [46], considering 27 group attributes results in over 7 billion intersectional groups, requiring a number of samples on the order of the human population to measure representation effectively. This makes it impractical to accurately measure the degree of unfairness at the finest granularity of intersectionality. In such cases, MPR can navigate the intersectionality-complexity tradeoff by using simpler function classes, like linear classifiers or low-depth decision trees. A more detailed mathematical discussion on the impact of function complexity on MPR is provided in Section 3.

Remark 1. A main challenge in the MPR framework is determining the reference distribution R. Ideally, R should be well-balanced across intersectional groups, representing a diverse population. Since such complex distributions rarely have been known in straightforward analytic forms, we approximate R by randomly sampling from a curated dataset $\mathcal{D}$ that has ideal representation statistics. An example of $\mathcal{D}$ is the FairFace dataset [30], which is specifically designed to be balanced across race, gender, and age groups. We could also imagine sampling individuals according to the Census data of a given country, such as US Bureau of Labor Statistics [6] data, as well as demographic information from other relevant international institutes and entities, such as the United Nations Population Division [1], the World Bank [7], Eurostat [3], ILOSTAT [4], or the Organisation for Economic Co-operation and Development (OECD) [5]. In Section 4, we discuss how we select $\mathcal{D}$ in our experiments.

3.1. MPR with Empirical Distributions

We approximate MPR in (1) with empirical distributions, by generating k images from G_q, and sampling m images from a reference dataset with distribution R:

$$\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) = \sup_{c\in\mathcal{C}} \left| \frac{1}{k}\sum_{i=1}^{k} c(x_i^g) - \frac{1}{m}\sum_{j=1}^{m} c(x_j^r) \right| \quad (2)$$

where $\widehat{G}_q = \{x_i^g\}_{i=1}^k$ and $\widehat{R} = \{x_j^r\}_{j=1}^m$ are the generated and reference samples used to approximate G_q and R. To quantify the closeness between the true MPR defined in (1) and its empirical approximation in (2), we derive bounds for the approximation error in Proposition 1.

Proposition 1 (Generalization Gap of MPR)**.** For any given prompt q, let $\widehat{G}_q$ and $\widehat{R}$ be the sets of generated and reference samples used to approximate G_q and R, respectively. Then, for any bounded class of functions $\mathcal{C}$ and any $\delta > 0$,

$$\left|\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) - \mathrm{MPR}(\mathcal{C}, G_q, R)\right| \leq 2\mathcal{R}_{\widehat{G}_q}(\mathcal{C}) + 2\mathcal{R}_{\widehat{R}}(\mathcal{C}) + B\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2(k+m)}} \quad (3)$$

with probability at least $1-\delta$, where $B = \sup_{\substack{c\in\mathcal{C} \\ X\neq X'}} |c(X) - c(X')|$ and $\mathcal{R}_{\mathcal{A}}(\mathcal{C})$ is the Rademacher complexity of $\mathcal{C}$ relative to the set $\mathcal{A}$.

Proposition 1 reveals the fundamental trade-off among three key factors: the approximation error, the sample sizes of $\widehat{G}_q$ and $\widehat{R}$, and the complexity of function class $\mathcal{C}$. The approximation error decreases as we reduce the complexity of $\mathcal{C}$ and increase the sample sizes of $\widehat{G}_q$ and $\widehat{R}$. While Proposition 1 offers guidance for selecting appropriate sample sizes to achieve a target error threshold with a fixed $\mathcal{C}$, we extend our analysis in two directions. First, we present empirical results in Section D.2 that demonstrate how the approximation error varies with the complexity of $\mathcal{C}$. Second, we discuss practical heuristics for optimal selection of k and m in Section E. Beyond representational harm related to gender, age, and race, there are also unfairness tied to factors like disabilities, emotions, and image-backgrounds, as qualitatively discussed by [11, 42]. Increasing the complexity of function class $\mathcal{C}$ enables MPR to capture and measure these subtle representational disparities with greater precision. Guided by Proposition 1, we can methodically determine the sample sizes needed to achieve desired measurement accuracy for a given complexity of $\mathcal{C}$. This approach allows MPR to reliably quantify both conventional demographic and more nuanced representational disparities.

Thus far, we have considered MPR for a single prompt-conditioned generative distribution, G_q. Next, we approximate the average MPR of the T2I generation model over the entire prompt space by leveraging an empirical distribution of sampled prompts. Proposition 2 provides a bound on the difference between the empirical mean of MPR across N prompts and the true expected MPR, ensuring that MPR generalizes to new, unseen prompts.

Proposition 2 (Generalization of MPR over all Prompts)**.** Let P denote the distribution of prompts, and $\{q_1, \ldots, q_N\}$ denote a set of independent prompts sampled from P. Then,

for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\text{MPR}(\mathcal{C},\widehat{G}_{q_i},\widehat{R}) - \mathbb{E}_{Q\sim P}[\text{MPR}(\mathcal{C},G_Q,R)]\right| \geq \epsilon\right)$$
$$\leq \exp\left(-\frac{\epsilon^2 N}{8}\right) + \exp\left(-\frac{2(k+m)}{B^2}\left(\frac{\epsilon}{2} - 2\lambda\right)\right) \quad (4)$$

where k and m are the numbers of generated and reference samples used in the calculation of $\text{MPR}(\mathcal{C}, \widehat{G}_{q_i}, \widehat{R})$, Q is the random variable representing a prompt, $Q \sim P$, and $\lambda = \sup_{Q\sim P} \mathcal{R}_{\widehat{G}_Q}(\mathcal{C}) + \mathcal{R}_{\widehat{R}}(\mathcal{C})$.

3.2. MPR computation according to $\mathcal{C}$

In this paper, we consider two main classes of functions to describe sensitive intersectional groups: 1) bounded linear functions and 2) binary decision trees. The following results characterize MPR for each function class.

Proposition 3. For the class of bounded linear functions, $\mathcal{C} = \{c : w^T x \big| w \in \mathbb{R}^d, \|w\| \leq 1\}$,

$$\text{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) = \frac{1}{||\tilde{a}^T X||}\tilde{a}^T X X^T \tilde{a}, \quad (5)$$

in which, $\tilde{a} \in \mathbb{R}^{k+m}$ has i-th entry given by $\tilde{a}_i = \mathbb{1}_{i\leq k}\frac{1}{k} - \mathbb{1}_{i>m}\frac{1}{m}$ and $X \in \mathbb{R}^{(k+m)\times d}$ is a matrix with its first k rows representing the generated images $\{x_i^g\}_{i=1}^k$ and the last m rows representing the reference images sampled $\{x_i^r\}_{i=1}^m$.

Next, we present the closed form of MPR with binary decision trees. In this case, we assume $x_i^g, x_i^r \in \{-1, +1\}^n$ to be feature vectors denoting the respective sample's group membership (binary) in n sensitive attributes. For example, if $n = 2$ with the corresponding sensitive attributes being race and gender, an image of a person belonging to race 1 and gender 1 is denoted by the vector $x = [+1, +1]$.

Proposition 4. For the class of binary decision trees of depth $\ell \leq n$, where $\mathcal{C} = \{c : \{-1, +1\}^n \to \{-1, +1\}\}$,

$$\text{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) = \max_{I\subset\{1,2,\ldots,n\}, |I|=\ell} 2\text{TV}(\tilde{G}_q^I, \tilde{R}^I) \quad (6)$$

where $\tilde{G}_q^I$ and $\tilde{R}^I$ are the marginal distributions corresponding to $\widehat{G}_q$ and $\widehat{R}$ over the attributes in set I, and $\text{TV}(A, B)$ denotes the total variation between distributions A and B.

4. Quantitative Assessment of Representational Harm Using MPR

We investigate representational harms inherent in various T2I models using the MPR metric. Unlike previous studies, e.g., [16, 23, 59], which primarily focused on single-group such as occupational stereotypes, we extend the analysis to multiple demographic groups and quantify representational harms related to traits and disabilities, which were qualitatively studied in the previous works [11, 42].

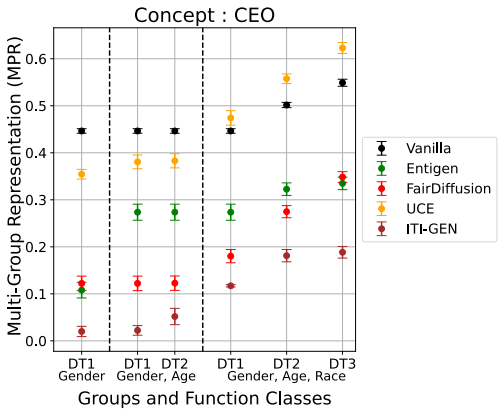


Figure 2. **MPRs according to C and group attributes when applying existing bsaeline methods to mitigate unfairness of SD v1.4 for "CEO".** DT represents decision trees, and the number following it indicates its depth. The text under DT refers to the group attributes used.

Experimental setup. For each concept under investigation, we generate a set of images using prompts such as *"a portrait photo of a {concept}"*. The MPR metric is then calculated to assess how biased a T2I model's representation is relative to a reference dataset in terms of multiple groups. We generate 1,000 images for each experiment and use bootstrapping with 1,000 resamples with replacement, performed 100 times, to compute the average MPR and its standard deviation. Unless specified otherwise, we use the Fairface [30] test dataset as our reference dataset, which is carefully designed to be balanced for gender, age, and race. The group labels of images for gender, age, and race are obtained by linear classifiers probed on CLIP embedding space with Fairface[2]. When using decision trees as $\mathcal{C}$, the outputs of these classifiers are encoded as one-hot vectors, i.e., binary vectors where all elements are 0 except for one element, which is 1. We utilize three open-source T2I models, including Stable Diffusion v1.4 [58], Stable Diffusion v2.1 [62], and SDXL [54]. We report MPR values for several intervention methods: the baseline models without any intervention (denoted here as Vanilla), Entigen [9], FairDiffusion [23], ITI-GEN [66], UCE [24]. Unless otherwise specified, all baselines are adjusted to mitigate representational harm based on the reference used by MPR. UCE, Entigen, and FairDiffusion are implemented to address the unfairness only for gender, while ITI-GEN considers intersectional groups regarding gender, age, and race. The other implementation details are in Section C.

[2]We note a body of work discussing the ethics of predicting sensitive attributes [27], but do not presume to address this issue here.

Table 1. **MPR results for several traits**. Every MPR is obtained with groups of "Gender", "Age" and "Race" and decision trees. The splits highlighted in bold font indicate those used in the decision trees with a depth of 1. The number in parenthesis represents the standard deviation estimated from bootstrapping the generated images.

		Attractive	Emotional	Exotic	Poor	Terrorist	Thug
SD v1.4	DT($d=1$)	0.67 ($\pm$0.01)	0.46 ($\pm$0.01)	0.39 ($\pm$0.01)	0.53 ($\pm$0.01)	0.47 ($\pm$0.00)	0.81 ($\pm$0.01)
	DT($d=3$)	0.72 ($\pm$0.01)	0.62 ($\pm$0.01)	0.59 ($\pm$0.01)	0.55 ($\pm$0.01)	0.61 ($\pm$0.00)	0.84 ($\pm$0.01)
SD v2.1	DT($d=1$)	0.66 ($\pm$0.01)	0.52 ($\pm$0.01)	0.30 ($\pm$0.01)	0.50 ($\pm$0.00)	0.48 ($\pm$0.01)	0.47 ($\pm$0.01)
	DT($d=3$)	0.71 ($\pm$0.01)	0.53 ($\pm$0.01)	0.41 ($\pm$0.02)	0.60 ($\pm$0.01)	0.79 ($\pm$0.01)	0.60 ($\pm$0.01)
SDXL	DT($d=1$)	0.73 ($\pm$0.01)	0.51 ($\pm$0.02)	0.43 ($\pm$0.01)	0.53 ($\pm$0.01)	0.47 ($\pm$0.00)	0.72 ($\pm$0.01)
	DT($d=3$)	0.77 ($\pm$0.01)	0.55 ($\pm$0.01)	0.56 ($\pm$0.01)	0.56 ($\pm$0.01)	0.79 ($\pm$0.01)	0.80 ($\pm$0.01)
Splits of DT($d=3$) SD v1.4		Old Southeast Asian **White**	**Male** White Latino	Old Black **Indian**	Male **Old** **Indian**	**Male** Black East Asian	Male Old **Black**
Splits of DT($d=3$) SD v2.1		Male Old **White**	Old East Asian **White**	Old **Indian** Latino	**Old** Indian White	Male Indian **Middle Eastern**	**Male** Black Latino
Splits of DT($d=3$) SDXL		Male Old **White**	Old East Asian **White**	**Male** Age White	Male **Old** Black	**Male** Indian Middle Eastern	Male Old **Black**

4.1. Application of MPR across diverse scenarios

MPRs for profession representation. Our MPR framework can adapt to a variety of scenarios, which differ in which groups are focused on and how fine-grained intersectional groups should be considered. To emulate these situations, we apply MPR while increasing the number of groups considered – starting with gender and then adding age and race – and simultaneously increasing the depth of the decision tree classifiers.

Fig. 2 shows the MPR values for images generated by the vanilla SD v1.4 model and baseline methods for the prompts using "CEO". From the figure, we observe that as the number of group attributes and the depth of the decision trees (i.e., the complexity of the function class $\mathcal{C}$) increase, the MPR value increases. This is because MPR identifies the most over- or under-represented group from the reference population over larger C. MPR exposes critical limitations in existing harm mitigation methods that were specifically developed to address gender disparity. While FairDiffusion, UCE, and Entigen demonstrate effective disparity reduction in their intended domain of gender fairness, their performance deteriorates significantly when considering intersectional attributes of age and race simultaneously. When considering race together, UCE's performance even falls below that of the vanilla model, highlighting the challenges of addressing multiple dimensions of representational harm beyond their original gender-focused design. These quantitative findings are visually corroborated by generated examples shown in Fig. 8 and 9. For a comprehensive analysis across a wider range of occupations, we present additional MPR results in Section 5.

MPRs for trait representation. MPR can calculate representational harm for any query, given the group estimation method for images. Specifically, we investigate the unfairness in intersectional groups for six human-related traits –"attractive", "emotional", "exotic", "poor", "terrorist" and "thug" – which were only qualitatively assessed in the previous research [11][3]. Table 1 quantitatively presents the extent of intersectional group unfairness for these traits using MPR values across the three T2I models, considering gender, age, and race groups. We also report the features used as splits in the decision trees selected to compute MPR, showing that using decision trees provides interpretable results. It shows that all three models exhibit large MPR values for each trait, showing a significant degree of unfairness. Importantly, as shown in the results for the "exotic" trait, *we can measure more severe unfairness when considering more fine-grained intersectional groups*. Finally, the splits used in the decision trees allow us to interpret which aspects of the T2I models exhibit the most unfairness. For example, in the cases of "attractive" and "terrorist", the splits mainly involve "White" and "Middle Eastern" groups, which align with the qualitative findings from [11].

Table 2. **MPRs for disability representation.** WC is short for a wheelchair user.

		MPR	Splits
SD v1.4	DT ($d=1$)	0.53 ($\pm$0.01)	White
	DT ($d=2$)	0.54 ($\pm$0.02)	WC / White
SD v2.1	DT ($d=1$)	0.50 ($\pm$0.01)	WC
	DT ($d=2$)	0.54 ($\pm$0.01)	WC / Asian
SDXL	DT ($d=1$)	0.55 ($\pm$0.01)	WC
	DT ($d=2$)	0.61 ($\pm$0.01)	WC / Black

[3]Here we used the same prompt of "the portrait photo of a {concept}" person as [11].

Table 3. **Query-Contextual MPR.** Balanced (left) and contextual (right) reference set. MPR is calculated with the decision trees with a depth of 1 and "gender" as a sensitive attribute.

Reference	Equal	Google Images
SD v1.4	0.40 ($\pm$ 0.01)	0.69 ($\pm$ 0.01)
FairDiffusion [23]	0.03 ($\pm$ 0.01)	0.32 ($\pm$ 0.01)

MPRs for disability representation. The MPR framework's flexibility in accommodating different group estimators and reference datasets enables us to quantify diverse forms of representational harm, extending beyond traditional demographic categories like gender, age, and race to include more nuanced social dimensions. A compelling example of such subtle harm emerges in the context of disability representation, which Mack et al. [42] explored in their analysis of images generated from the prompt *"a photo of a person with a disability"*. Their work revealed a concerning pattern where generated images disproportionately depicted wheelchair users while underrepresenting individuals with visual, auditory, or cognitive disabilities. Building upon their qualitative observations, we provide a qualitative analysis using MPR. Our study focuses on the overrepresentation of mobility disabilities and its intersection with racial disparity. We use reference statistics from the U.S. Centers for Disease Control and Prevention (CDC) [2], which provide detailed data on the intersectional distribution of race and disability types. For tractability, we group the disability into mobility disabilities (represented by wheelchair users) and non-mobility disabilities (encompassing cognitive, visual, and auditory impairments), and consider four major racial groups: White, Black, Asian, and Latino populations. To detect wheelchair presence in generated images, we employ the BLIP2 Visual Question Answering (VQA) system with the query *"What objects are in the image?"*, deriving pseudo-labels from its responses. Table 2 presents MPR values and their corresponding decision tree splits at depths 1 and 2. The analysis demonstrates that the three T2I models exhibit compounded bias: *not only do they disproportionately generate images of wheelchair users over other forms of disability, but they also show substantial racial skew within these representations*. This is visually substantiated in Figure 10, where the SD v1.4 model frequently depicts white wheelchair users.

4.2. Use cases unlocked by MPR

The formulation of MPR and choice of $\mathcal{C}$ allows MPR to be used in settings where prior fairness metrics are unable to measure bias or fail to accurately capture the nuances of representation. We demonstrate a use case for querying a reference dataset to handle specific social and historical contexts and then discuss the benefits of exploring a general class of functions $\mathcal{C}$.

MPR is prompt contextual. MPR's flexibility extends to evaluating prompt-specific representational requirements, as

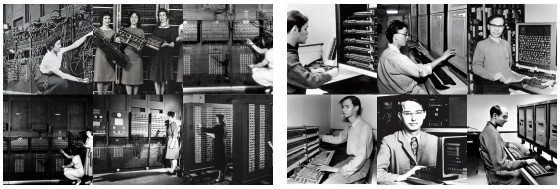


Figure 3. Left: A random sample of the top 100 Google Images results for *"A photo of a computer programmer for the ENIAC"*. Right: Sample generated images by SD 1.4 for the same prompt.

demonstrated through our analysis of the prompt *"A photo of a computer programmer for the ENIAC."* It is well-known that early computer programming pioneers were predominantly women, with six women programmers specifically responsible for the ENIAC [10, 35]. Given this historical context, accurate image generation should exclusively depict women programmers. To establish a reference distribution, we analyzed the top 100 Google image search results for this prompt, which mostly showed women programmers (Fig. 3, left). Table 3 presents MPR measurements comparing various model generations against two reference points: an equal gender distribution and the historically accurate women-only representation derived from our Google image search query. The results reveal two distinct representational harms: First, the vanilla model strongly over-represents male programmers (Fig. 3, right), diverging from both equal representation and historical accuracy. Second, while a fair generative baseline (i.e., FairDiffusion) achieve gender parity, they fall short of capturing the historically appropriate representation of exclusively women programmers. This analysis demonstrates MPR's capability to identify not only deviations from equal representation but also failures to preserve important historical context. The results of other baselines are described in Section D.3 in the supplementary material.

MPR can scale to a large set of attributes. Balancing representation for each identity combination leads to computational complexity that grows exponentially with the number of attribute categories to account for all possible intersectional groups. In contrast, MPR's function class $\mathcal{C}$ can efficiently process attribute vectors of arbitrary dimensionality, enabling the measurement of complex, high-dimensional representational patterns. This capability is particularly relevant given recent developments like the multifaceted identity embedding framework introduced in [60]. MPR can evaluate representational fairness over such rich embeddings without incurring additional computational overhead, though we note that the specific embeddings from [60] remain proprietary and were not available for our experimental validation.

5. Fine-tuning T2I models for MPR

We propose a fine-tuning method to reduce representational harm in T2I models in terms of MPR. Given a reference

dataset, we can improve MPR of a T2I model by utilizing Equation (2) as an objective function. However, minimizing this function directly via gradient descent is non-trivial for two main reasons: (1) estimating MPR requires generating a large number of images, which would result in significant computational overhead if done at every update, and (2) conducting gradient descent is challenging because of the supremum operation in (2). At each iteration, we can compute the c^* which maximizes MPR for a specific set of generated images using Props. 3 and 4. However, we observed that just updating the model to reduce MPR for the current c^* may not reduce it over the full set of functions (the supremum). For example, the model may first optimize gender representation, then switch to race representation, and in doing so lose the progress it gained on gender representation. To address the aforementioned issues, we store generated images and the c^* functions into buffers at each update. Then, we approximate MPR in every update step using the buffers together with the newly generated images and computed c^*. By doing so, we can avoid the need to generate a large number of images per iteration and address the supremum. The buffers are updated in a First-In-First-Out (FIFO) manner, constrained by predefined sizes. Additionally, minimizing only Equation (2) could cause the finetuned model θ_t at iteration t to deviate excessively from the pre-trained model θ_0, potentially degrading image generation quality. To prevent this, we incorporate regularization terms based on the CLIP and DINO models, as proposed in [59]. Our final objective function is $\mathcal{L}_{\text{obj}}(\theta_t) = \mathcal{L}_{\text{MPR}}(\theta_t) + \lambda \mathcal{L}_{\text{img}}(\theta_t)$, where

$$\mathcal{L}_{\text{MPR}}(\theta_t) = \sum_{c \in \hat{\mathcal{C}}} \Big| \frac{1}{k} \sum_{x_i \in \hat{G}} c(x_i) - \frac{1}{m} \sum_{x_j \in D} c(x_j) \Big|,$$

$$\mathcal{L}_{\text{img}}(\theta_t) = \frac{1}{N} \sum_{i=1}^{N} \Big[(1 - \cos(\text{CLIP}(x_i), \text{CLIP}(o_i)) + (1 - \cos(\text{DINO}(x_i), \text{DINO}(o_i))) \Big].$$

Here, λ is a hyperparameter for the regularization terms, x_i, o_i are samples generated from θ_t, θ_0 respectively, and $\hat{G}$ and $\hat{\mathcal{C}}$ contain generated images and c functions obtained from the current and previous updates. We adopt the gradient-adjusting technique [59] to prevent gradient explosion and utilize LoRA [29] for computationally efficient learning. Our fine-tuning method is detailed in Alg. 1.

Performance comparison. We validate the effectiveness of our method by examining its ability to improve MPR values for the occupation representation. Specifically, we applied our method and existing baseline methods for seven professions (chef, therapist, taxi driver, flight attendant, housekeeper, nurse) that were qualitatively assessed in [11]. We evaluated all methods in terms of representational harm and image quality using two types of MPR values and the CLIP similarity between the occupation prompts and the generated images. To ensure a fair comparison, all baseline methods were also adapted to use statistics derived from the reference dataset.

Table 4. **Performance comparison.** (G) and (GAR) indicate the group attributes considered: gender, or gender, age, and race, respectively. The reported MPR values and CLIP scores are averages across seven professions (chef, therapist, taxi driver, flight attendant, housekeeper, nurse).

	MPR (G)	MPR (GAR)	CLIP score
Vanilla	1.18 ($\pm$0.02)	1.89 ($\pm$0.02)	0.31 ($\pm$0.01)
Entigen [9]	1.05 ($\pm$0.03)	1.61 ($\pm$0.03)	0.30 ($\pm$0.01)
FairDiffusion [23]	0.27 ($\pm$0.04)	1.32 ($\pm$0.02)	0.31 ($\pm$0.02)
UCE [24]	0.80 ($\pm$0.05)	1.56 ($\pm$0.04)	0.31 ($\pm$0.01)
ITI-GEN [66]	0.13 ($\pm$0.04)	0.44 ($\pm$0.02)	0.28 ($\pm$0.03)
Ours	**0.08** ($\pm$0.05)	**0.25** ($\pm$0.03)	0.30 ($\pm$0.01)

Table 4 demonstrates that methods like FairDiffusion and UCE, which focus solely on a single group attribute (gender), perform poorly when the MPR metric considers multiple attributes (gender, age, and race). Although ITI-GEN achieves reasonable performance on both MPR metrics, its lack of consideration for image quality results in slightly degraded CLIP scores. In contrast, our method achieves the best performance across both types of MPR while maintaining robust CLIP scores, ensuring high image quality. The details of our method implementation are provided in Section C in the supplementary material.

6. Concluding Remarks and Limitations

We introduced Multi-Group Proportional Representation (MPR) as a comprehensive framework for measuring representational harm in text-to-image generative models. Unlike existing methods that often focus on a single or limited number of sensitive attributes, MPR supports harm measurement with respect to arbitrary sensitive attributes and their intersections. Through extensive experiments, we quantitatively measured significant representational disparities across professions, traits, and disabilities that prior works addressed only qualitatively. Additionally, we demonstrate that using MPR as an optimization objective during model fine-tuning effectively reduces intersectional disparities while maintaining image quality. While important challenges remain, including the need for robust attribute estimation methods and context-appropriate reference distributions, our work provides a systematic approach to measuring and mitigating representational harms. As text-to-image models become increasingly prevalent, frameworks like MPR will be crucial to ensuring that these systems serve and represent all members of society equitably.

Acknowledgments

We would like to thank Himabindu Lakkaraju for her helpful discussions and valuable insights during the preparation of this paper.

This work was supported in part by the National Research Foundation of Korea (NRF) grant [RS-2021-NR059237] and by Institute of Information & communications Technology Planning & Evaluation (IITP) grants [RS-2021-II211343, RS-2021-II212068, RS-2022-II220113, RS-2022-II220959] funded by the Korean government (MSIT). It was also supported by AOARD Grant No. FA2386-25-1-4015, Brain Korea 21 Plus Project and Hyundai Motor Chung Mong-Koo Foundation. AO is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-2140743. This material is based upon work supported by the National Science Foundation under awards CIF-2231707, CIF-2312667, and FAI-2040880.

References

[1] United nations department of economic and social affairs: Population division. https://population.un.org/wpp/. Accessed: 2024-11-14. 4

[2] Centers for disease control and prevention. https://dhds.cdc.gov/SP. Accessed Sep 11, 2024. 7

[3] Eurostats data. https://ec.europa.eu/eurostat/web/population-demography/demography-population-stock-balance/database. Accessed: 2024-11-14. 4

[4] International labour organization statistics. https://ilostat.ilo.org/. Accessed: 2024-11-14. 4

[5] Oecd data. https://www.oecd.org/en/data.html. Accessed: 2024-11-14. 4

[6] Us bureau of labor statistics. employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity. https://www.bls.gov/cps/cpsaat11.htm. Accessed: 2024-11-09. 4

[7] World bank open data. https://data.worldbank.org/. Accessed: 2024-11-14. 4

[8] Adobe. Firefly, 2024. Accessed Sep 7, 2024. https://www.adobe.com/products/firefly/features/text-to-image.html. 1

[9] 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, 5, 8

[10] Jean Jennings Bartik. *Pioneer programmer: Jean Jennings Bartik and the computer that changed the world*. Truman State University Press, 2013. 7

[11] 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 *ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023. 1, 2, 3, 4, 5, 6, 8

[12] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 3

[13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities: A nonasymptotic theory of independence. 2013. 2

[14] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. 7

[15] Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. Tibet: Identifying and evaluating biases in text-to-image generative models. In *European Conference on Computer Vision*, pages 429–446. Springer, 2024. 3

[16] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *European Conference on Computer Vision*, pages 3043–3054, 2023. 1, 2, 3, 5

[17] Yujin Choi, Jinseong Park, Hoki Kim, Jaewook Lee, and Saerom Park. Fair sampling in diffusion models through switching mechanism. In *AAAI Conference on Artificial Intelligence*, pages 21995–22003, 2024. 3

[18] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 3

[19] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12225–12235, 2024. 1

[20] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12225–12235, 2024. 3

[21] Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems. *IEEE/CVF International Conference on Computer Vision*, 2023. 1, 3

[22] Kathleen C Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified? In *The AAAI-23 Workshop on Creative AI Across Modalities*, 2023. 2

[23] 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. *arXiv preprint arXiv:2302.10893*, 2023. 1, 2, 3, 5, 7, 8

[24] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *IEEE/CVF Winter Conference on Ap-*

plications of Computer Vision, pages 5111–5120, 2024. 3, 5, 8

[25] Google. Introducing gemini: our largest and most capable ai model, 2023. Accessed Sep 7, 2024. https://blog.google/technology/ai/google-gemini-ai/. 1

[26] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 3

[27] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *ACM Conference on Human Factors in Computing Systems*, pages 1–13, 2018. 5

[28] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948, 2018. 3

[29] 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. 2021. 8

[30] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 4, 5

[31] Eunji Kim, Siwon Kim, Chaehun Shin, and Sungroh Yoon. De-stereotyping text-to-image models through prompt tuning. *Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning*, 2023. 2, 3

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

[33] Yeongmin Kim, Byeonghu Na, Minsang Park, JoonHo Jang, Dongjun Kim, Wanmo Kang, and Il-Chul Moon. Training unbiased diffusion models from biased dataset. *International Conference on Learning Representations*, 2024. 3

[34] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 6

[35] Kathy Kleiman. *Proving ground: The untold story of the six women who programmed the world's first modern computer*. Hurst Publishers, 2022. 7

[36] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 7

[37] Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. *arXiv preprint arXiv:2311.17695*, 2023. 3

[38] Shuangqi Li, Chen Liu, Tong Zhang, Hieu Le, Sabine Süsstrunk, and Mathieu Salzmann. Controlling the fidelity and diversity of deep generative models via pseudo density. *Transactions on Machine Learning Research*, 2024. 3

[39] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[40] Hanjun Luo, Haoyu Huang, Ziye Deng, Xuecheng Liu, Ruizhe Chen, and Zuozhu Liu. Bigbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal llm. *arXiv preprint arXiv:2407.15240*, 2024. 1, 3

[41] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 7

[42] Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K Kane, and Cynthia L Bennett. "they only care to show us the wheelchair": disability representation in text-to-image ai models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2024. 1, 2, 4, 5, 7

[43] Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989. 2

[44] Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024. 3

[45] Midjourney. Midjourney.com, 2024. Accessed Sep 7, 2024. https://www.midjourney.com/home. 1

[46] Lucas Monteiro Paes, Carol Long, Berk Ustun, and Flavio Calmon. On the epistemic limits of personalized prediction. *Advances in Neural Information Processing Systems*, 35:1979–1991, 2022. 4

[47] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29 (2):429–443, 1997. 3

[48] Alex Oesterling, Claudio Mayrink Verdun, Alexander Glynn, Carol Xuan Long, Lucas Monteiro Paes, Sajani Vithana, Martina Cardone, and Flavio Calmon. Multi-group proportional representation in retrieval. In *Advances in Neural Information Processing Systems*, 2024. 2, 3

[49] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, pages 7053–7061, 2023. 2, 3

[50] Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness. In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 496–511, 2023. 3

[51] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6668–6678, 2024. 3

[52] Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023. 7

[53] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 3

[54] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach.

Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. 5

[55] Prabhakar Raghavan. Gemini image generation got it wrong. we'll do better. https://blog.google/products/gemini/gemini-image-generation-issue/. Accessed: 2024-11-09. 1

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. PMLR, 2021. 1

[57] Adi Robertson. Google apologizes for 'missing the mark' after gemini generated racially diverse nazis. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical. Accessed: 2024-11-09. 1, 3

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 5

[59] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. *arXiv preprint arXiv:2311.07604*, 2023. 1, 2, 3, 5, 8

[60] Hansa Srinivasan, Candice Schumann, Aradhana Sinha, David Madras, Gbolahan Oluwafemi Olanubi, Alex Beutel, Susanna Ricco, and Jilin Chen. Generalized people diversity: Learning a human perception-aligned diversity representation for people images. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 797–821, 2024. 7

[61] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. 3

[62] Stabel Diffusion 2.1. Stable diffusion v2.1 and dreamstudio updates 7-dec 22, 2022. Accessed September 10, 2024. https://stability.ai/news/stablediffusion2-1-release7-dec-2022. 5

[63] Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*, 2024. 1

[64] Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[65] Hidir Yesiltepe, Kiymet Akdemir, and Pinar Yanardag. Mist: Mitigating intersectional bias with disentangled cross-attention editing in text-to-image diffusion models. *arXiv preprint arXiv:2403.19738*, 2024. 1, 3

[66] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Itigen: Inclusive text-to-image generation. In *IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023. 1, 2, 3, 5, 8

[67] Eric Zhao, De-An Huang, Hao Liu, Zhiding Yu, Anqi Liu, Olga Russakovsky, and Anima Anandkumar. Scaling fair learning to hundreds of intersectional groups. 2022. 3