Supplementary Material

In this supplementary material, we provide detailed proofs, additional experimental results, and implementation details that complement our main paper. The appendix is organized as follows:

Section A outlines the limitations of our approach, discussing challenges related to attribute estimation, reference population collection, and prompt construction. Section B presents complete mathematical proofs for all propositions stated in Section 3 of the main paper. In particular, we establish the Generalization bound for the Multi-group proportional metric, stated in Proposition 1, and we characterize MPR for the class of bounded linear functions and decision trees, stated in Propositions 3 and 4 respectively. Section C provides comprehensive implementation details, including our methodology for group label creation and the implementation specifics of baseline and fine-tuning methods. Section D contains additional experimental results, including

i. MPR results for trait representation across multiple diffusion models.
ii. Analysis of empirical vs. true MPR gaps across different function classes.
iii. Additional baseline method results for the ENIAC programmer case.
iv. Supplementary qualitative results.

Section E discusses practical considerations for selecting parameters k and m in the MPR framework, including guidance for real-world applications and empirical validation.

A. Limitations

While MPR represents a significant advance in measuring and mitigating representational bias in text-to-image systems, there are several important limitations to our approach that warrant careful consideration. One fundamental challenge lies in our reliance on automated systems to estimate demographic attributes such as gender, race, and disability status. These estimation methods may not only perpetuate harmful categorization practices but could also contain inherent biases themselves. The binary classification of complex social identities oversimplifies human diversity and risks reinforcing problematic societal categorizations. This is particularly concerning for attributes like disability status, which are both technically challenging to detect and ethically sensitive to classify.

The selection of appropriate reference distributions presents another significant challenge. Determining what constitutes appropriate representation is inherently complex and context-dependent, with different stakeholders potentially holding varying views on fair representation. Historical data used for reference distributions may contain existing societal biases, creating a tension between reflecting historical accuracy and promoting more equitable representation. Additionally, there is a notable scarcity of balanced, intersectional datasets across many domains, limiting our ability to establish comprehensive benchmarks for fair representation. We note that FairFace dataset is one of the most comprehensive facial image datasets, which is debiased toward gender, age and race. While some of our MPR evaluations rely on the FairFace dataset along gender, age, and race axes, they are merely examples of use cases for MPR and do not imply a limitation of MPR. Furthermore, we believe that the cost and effort required to create robust reference distributions are ultimately inevitable for measuring and mitigating biases in image generation models. The key contribution of our work lies in proposing a flexible measurement framework that can adapt to such advanced reference distributions as they are developed.

Our current evaluation methodology also has limitations in terms of prompt engineering and assessment. The focus on simple prompt structures (e.g., "*a photo of concept*") may not adequately reflect the complexity and variety of real-world usage. More diverse prompt templates are needed to better represent natural language variation and capture cultural or contextual nuances. The challenge of evaluating abstract concepts and scaling across multiple languages and cultural contexts remains significant.

From a methodological perspective, we face several technical constraints. There exists a fundamental trade-off between function class complexity and sample size requirements, affecting the practical applicability of our approach. The computational costs of evaluating multiple intersectional categories simultaneously can be substantial, and maintaining image quality while optimizing for representational fairness presents ongoing challenges. Our current framework also has limited ability to capture temporal or contextual aspects of representation.

Looking toward future work, these limitations highlight several critical directions for research. We need to develop more nuanced and ethical approaches to attribute detection, create context-aware reference distribution frameworks, and expand to more complex prompt structures and evaluation scenarios. Improving computational efficiency and deepening engagement with affected communities will be crucial. Furthermore, integration with other fairness metrics and evaluation frameworks could provide more comprehensive assessments of representational fairness.

These challenges underscore the complexity of measuring and promoting fairness in generative AI systems. As these technologies continue to evolve and their societal impact grows, addressing these limitations will be essential for developing more equitable and responsible AI systems that truly serve all members of society.

B. Proofs

Proposition 1 restated: 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 $\delta > 0$,

$$\left|\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) - \mathrm{MPR}(\mathcal{C}, G_q, R)\right| \leq 2\mathcal{R}_{\widehat{G}_q}(\mathcal{C}) + 2\mathcal{R}_{\widehat{R}}(\mathcal{C}) + B\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2(k+m)}} \tag{7}$$

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 $\mathcal{A}$.

Proof. Define $f(c) := \mathrm{MPR}(\mathcal{C}, G_q, R) = \mathbb{E}_{G_q}[c(X_g)] - \mathbb{E}_R[c(X_r)]$ and $\hat{f}(c) := \mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) = \frac{1}{k}\sum_{i=1}^{k} c(x_i^g) - \frac{1}{m}\sum_{j=1}^{m} c(x_j^r)$. To start, note that,

$$\left|\sup_{c\in\mathcal{C}} \hat{f}(c) - \sup_{c\in\mathcal{C}} f(c)\right| \leq \sup_{c\in\mathcal{C}} \left|\hat{f}(c) - f(c)\right| \tag{8}$$

which implies,

$$\begin{aligned}
\mathbb{P}\Big(\left|\sup_{c\in\mathcal{C}} \hat{f}(c) - \sup_{c\in\mathcal{C}} f(c)\right| \geq \epsilon\Big) &\leq \mathbb{P}\Big(\sup_{c\in\mathcal{C}} \left|\hat{f}(c) - f(c)\right| \geq \epsilon\Big) &(9)\\
&\leq \mathbb{P}\Big(\sup_{c\in\mathcal{C}} \Big(\hat{f}(c) - f(c)\Big) \geq \frac{\epsilon}{2}\Big) + \mathbb{P}\Big(\sup_{c\in\mathcal{C}} \Big(f(c) - \hat{f}(c)\Big) \geq \frac{\epsilon}{2}\Big) &(10)
\end{aligned}$$

Next, we analyze each probability term in (10). Define $g(x_1^g, \ldots, x_k^g, x_1^r, \ldots, x_m^r) = \sup_{c\in\mathcal{C}}(\hat{f}(c) - f(c))$. Let $\{z\}_{i=1}^{k+m} = \{\{x_i^g\}_{i=1}^n, \{x_i^r\}_{i=1}^m\}$ be the concatenation of $\mathcal{G}$ and $\mathcal{D}$. For any $i \in \{1, \ldots, k+m\}$,

$$\begin{aligned}
\left|g(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_{k+m}) - g(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_{k+m})\right| &\leq \sup_{c\in\mathcal{C}} \frac{1}{\min\{m,k\}}\left|c(z_i) - c(z_i')\right| \\
&= \frac{B}{\min\{m,k\}}, &(11)
\end{aligned}$$

where (11) follows from (8). Then, from McDiarmid's inequality [13, 43], we have,

$$\mathbb{P}\Big(\sup_{c\in\mathcal{C}}(\hat{f}(c) - f(c)) \geq \epsilon + \mathbb{E}[\sup_{c\in\mathcal{C}}(\hat{f}(c) - f(c))]\Big) \leq e^{-\frac{2\epsilon^2(m+k)}{B^2}}.$$

Now, we bound the expectation of the right-hand side inside the probability with a symmetrization argument. Indeed, Let

σ_i be independent Rademacher random variables (uniformly distributed on $\{-1, 1\}$.

$$\mathbb{E}\big[\sup_{c\in\mathcal{C}}(\hat{f}(c) - f(c))\big] = \mathbb{E}\left[\sup_{c\in\mathcal{C}}\Big(\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) - \mathbb{E}\big[\mathrm{MPR}(\mathcal{C}, \widehat{G}'_q, \widehat{R}')\big]\Big)\right] \tag{12}$$

$$= \mathbb{E}\left[\sup_{c\in\mathcal{C}}\Big(\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}\Big) - \mathbb{E}\big[\mathrm{MPR}(\mathcal{C}, \widehat{G}'_q, \widehat{R}')|\widehat{G}_q, \widehat{R}\big]\right] \tag{13}$$

$$= \mathbb{E}\left[\sup_{c\in\mathcal{C}}\mathbb{E}\big[\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) - \mathrm{MPR}(\mathcal{C}, \widehat{G}'_q, \widehat{R}')|\widehat{G}_q, \widehat{R}\big]\right] \tag{14}$$

$$\leq \mathbb{E}\left[\mathbb{E}\big[\sup_{c\in\mathcal{C}}(\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) - \mathrm{MPR}(\mathcal{C}, \widehat{G}'_q, \widehat{R}'))|\widehat{G}_q, \widehat{R}\big]\right] \tag{15}$$

$$= \mathbb{E}\left[\sup_{c\in\mathcal{C}}\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) - \mathrm{MPR}(\mathcal{C}, \widehat{G}'_q, \widehat{R}')\right] \tag{16}$$

$$= \mathbb{E}\sup_{c\in\mathcal{C}}\left[\frac{1}{k}\sum_{i=1}^{k}c(x_i) - \frac{1}{m}\sum_{i=1}^{m}c(x_i) - \frac{1}{k}\sum_{i=1}^{k}c(\tilde{x}_i) + \frac{1}{m}\sum_{i=1}^{m}c(\tilde{x}_i)\right] \tag{17}$$

$$= \mathbb{E}\sup_{c\in\mathcal{C}}\left[\frac{1}{k}\sum_{i=1}^{k}\sigma(c(x_i) - c(\tilde{x}_i)) - \frac{1}{m}\sum_{i=k+1}^{k+m}\sigma(c(x_i) - c(\tilde{x}_i))\right] \tag{18}$$

$$\leq \mathbb{E}\left[\sup_{c\in\mathcal{C}}\frac{1}{k}\sum_{i=1}^{k}\sigma c(x_i) + \sup_{c\in\mathcal{C}}\frac{1}{k}\sum_{i=1}^{k}\sigma c(\tilde{x}_i) + \sup_{c\in\mathcal{C}}\sum_{i=k+1}^{k+m} -\sigma c(\tilde{x}_i) + \sup_{c\in\mathcal{C}}\sum_{i=k+1}^{k+m} -\sigma c(\tilde{x}_i)\right] \tag{19}$$

$$= 2\mathbb{E}\sup_{c\in\mathcal{C}}\frac{1}{k}\sum_{i=1}^{k}\sigma c(x_i) + 2\mathbb{E}\sup_{c\in\mathcal{C}}\frac{1}{k}\sum_{i=n+1}^{k+m}\sigma c(x_i) \tag{20}$$

$$= 2\mathcal{R}_{\widehat{\mathcal{G}}_q}(\mathcal{C}) + 2\mathcal{R}_{\widehat{R}}(\mathcal{C}). \tag{21}$$

where, by an abuse of notation, we denote by $\widehat{G}'_q, \widehat{R}'$ the symmetric i.i.d. copies of G_q, R and $\mathcal{G}'_q$ and $\mathcal{R}'$ denote the sample sets (with samples $\tilde{x}_i$) obtained by the symmetric random variables which empirical distributions given by $\widehat{G}'_q$ and $\widehat{R}'$, respectively.

Now, by choosing $\frac{\epsilon}{2} = t + 2\mathcal{R}_{\hat{\mathcal{G}}_q}(\mathcal{C}) + 2\mathcal{R}_{\hat{\mathcal{R}}}(\mathcal{C})$ and $t = B\sqrt{\frac{\ln\frac{1}{\delta}}{2(k+m)}}$, from (10), we have,

$$\mathbb{P}\Big(\Big|\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) - \mathrm{MPR}(\mathcal{C}, G_q, R)\Big| \geq \epsilon\Big) \leq 2e^{-\frac{2\epsilon^2(k+m)}{B^2}}. \tag{22}$$

By taking $1 - \delta = 1 - 2e^{-\frac{2\epsilon^2(k+m)}{B^2}}$, we prove that,

$$\Big|\mathrm{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) - \mathrm{MPR}(\mathcal{C}, G_q, R)\Big| \leq 2\mathcal{R}_{\hat{\mathcal{G}}_q}(\mathcal{C}) + 2\mathcal{R}_{\hat{\mathcal{R}}}(\mathcal{C}) + B\sqrt{\frac{\ln\frac{1}{\delta}}{2(k+m)}} \tag{23}$$

with probability at least $1 - \delta$. □

Proposition 2 restated: Let P denote the distribution of prompts, and $\{q_1, \ldots, q_N\}$ denote a set of independent prompts sampled from P. Then,

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

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

Proof. First, we decompose the difference using the triangle inequality:

$$\left|\frac{1}{N}\sum_{i=1}^{N}\text{MPR}(C,\hat{G}_{q_i},\hat{R}_{q_i})-\mathbb{E}_{Q\sim P}[\text{MPR}(C,G_Q,R_Q)]\right| \leq \left|\frac{1}{N}\sum_{i=1}^{N}\text{MPR}(C,\hat{G}_{q_i},\hat{R}_{q_i})-\frac{1}{N}\sum_{i=1}^{N}\text{MPR}(C,G_{q_i},R_{q_i})\right| + \left|\frac{1}{N}\sum_{i=1}^{N}\text{MPR}(C,G_{q_i},R_{q_i})-\mathbb{E}_{Q\sim P}[\text{MPR}(C,G_Q,R_Q)]\right|$$

We start by bounding the first term. From Proposition 1, for any fixed q:

$$\mathbb{P}(|\text{MPR}(C,\hat{G}_q,\hat{R}_q)-\text{MPR}(C,G_q,R_q)|\geq\epsilon/2)\leq\exp\left(-\frac{2(m+n)(\epsilon/2-2\lambda)}{B^2}\right)$$

where $\lambda=\sup_{Q\sim P}(R_{G_Q}(C)+R_{R_Q}(C))$. Therefore:

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}[\text{MPR}(C,\hat{G}_{q_i},\hat{R}_{q_i})-\text{MPR}(C,G_{q_i},R_{q_i})]\right|\geq\epsilon/2\right)\leq\exp\left(-\frac{2(m+n)(\epsilon/2-2\lambda)}{B^2}\right)$$

Now, for the second term, we use Hoeffding's inequality since $\text{MPR}(C,G_q,R_q)$ is bounded in $[0,1]$ and the prompts are sampled i.i.d:

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\text{MPR}(C,G_{q_i},R_{q_i})-\mathbb{E}_{Q\sim P}[\text{MPR}(C,G_Q,R_Q)]\right|\geq\epsilon/2\right)\leq\exp\left(-\frac{N\epsilon^2}{8}\right)$$

To finish, by using a union bound:

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}\text{MPR}(C,\hat{G}_{q_i},\hat{R}_{q_i})-\mathbb{E}_{Q\sim P}[\text{MPR}(C,G_Q,R_Q)]\right|\geq\epsilon\right)&\leq\mathbb{P}(\text{First term}\geq\epsilon/2)+\mathbb{P}(\text{Second term}\geq\epsilon/2)\\&\leq\exp\left(-\frac{\epsilon^2N}{8}\right)+\exp\left(-\frac{2(m+n)(\epsilon/2-2\lambda)}{B^2}\right)\end{aligned}$$

□

Remark 2. The bound in Proposition 2 can be tightened if an estimate of the variance of $\text{MPR}(C,G_q,R)$ across prompts is available. By applying Bernstein's inequality instead of Hoeffding's inequality for the prompt sampling error, we obtain:

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N}\text{MPR}(C,\hat{G}_{q_i},\hat{R})-\mathbb{E}_{Q\sim P}[\text{MPR}(C,G_Q,R)]\right|\geq\varepsilon\right)\leq 2\exp\left(-\frac{N\varepsilon^2}{8\sigma^2+4B\varepsilon/3}\right) \tag{24}$$

$$+\,2\exp\left(-\frac{2(k+m)(\varepsilon/4-2\lambda)^2}{B^2}\right) \tag{25}$$

where $\sigma^2=\mathbb{E}_{Q\sim P}[(\text{MPR}(C,G_Q,R)-\mathbb{E}_{Q\sim P}[\text{MPR}(C,G_Q,R)])^2]$ is the variance of $\text{MPR}(C,G_q,R)$ across prompts sampled from P. The first term bounds the error from prompt sampling, while the second term bounds the average estimation error across prompts. This variance-aware bound becomes particularly valuable when σ^2 is small relative to the range B, which is often the case for specific prompt categories. In practical applications, σ^2 can be estimated empirically from the sampled prompts using $\hat{\sigma}^2=\frac{1}{N-1}\sum_{i=1}^{N}[\text{MPR}(C,\hat{G}_{q_i},\hat{R})-\frac{1}{N}\sum_{j=1}^{N}\text{MPR}(C,\hat{G}_{q_j},\hat{R})]^2$, allowing for adaptive confidence intervals that automatically tighten when MPR values exhibit low variance across the prompt distribution.

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

$$\text{MPR}(\mathcal{C},\widehat{P},\widehat{Q})=\frac{1}{||\tilde{a}^TX||}\tilde{a}^TXX^T\tilde{a}, \tag{26}$$

in which, $\tilde{a}\in\mathbb{R}^{k+m}$ has i-th entry given by $\tilde{a}_i=\mathbb{1}_{i\leq k}\frac{1}{k}-\mathbb{1}_{i>n}\frac{1}{m}$ and $X\in\mathbb{R}^{(k+m)\times d}$ is a matrix (row-wise) concatenated with a set of generated images and the reference dataset.

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

$$\text{MPR}(\mathcal{C}, \widehat{P}, \widehat{Q}) = \sup_{c \in \mathcal{C}} \Big| \frac{1}{k} \sum_{i \in \mathcal{G}} c(x_i^g) - \frac{1}{m} \sum_{j \in \mathcal{D}} c(x_j^r) \Big| \tag{27}$$

$$= \sup_{w : \|w\|^2 \leq 1} \tilde{a}^T X w \tag{28}$$

For a fixed X and $\tilde{a}$, $\tilde{a}^T X w$ is maximized when w is aligned along $\tilde{a}^T X$ with the maximum norm allowed. i.e.,

$$\arg \sup_{w : \|w\|^2 \leq 1} \tilde{a}^T X w = \frac{X^T \tilde{a}}{\|X^T \tilde{a}\|}. \tag{29}$$

Therefore, from (28),

$$\text{MPR}(\mathcal{C}, \widehat{P}, \widehat{Q}) = \sup_{w : \|w\|^2 \leq 1} \tilde{a}^T X w = \frac{\tilde{a}^T X X^T \tilde{a}}{\|\tilde{a}^T X\|}. \tag{30}$$

□

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

$$\text{MPR}(\mathcal{C}, \widehat{G}_q, \widehat{R}) = \max_{I \subset \{1,2,\ldots,n\}, |I| = \ell} 2\text{TV}(\tilde{G}_q^I, \tilde{R}^I) \tag{31}$$

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 distance between distributions A and B.

Proof. Let $c \in \mathcal{C} = DT^k$ be any decision tree of depth at most k, specified by a subset of of attributes $I \subset \{1, \ldots, d\}$. Each $c \in \mathcal{C}$ corresponds to a specific subset of attributes I and a set of group assignments $c_i \in \{-1, +1\}$ for $i = \{1, \ldots, 2^{|I|}\}$, as it partitions the dataset into $2^{|I|}$ disjoint sets $A_1^I, ..., A_{2^{|I|}}^I$, where $|I| = k$. On each A_i^I, c takes a constant value $c_i = \{-1, +1\}$. Then:

$$\sup_{c \in \mathcal{C}} |E_{G_q}[c(X_g)] - E_R[c(X_r)]| = \sup_{c \in \mathcal{C}} \left| \sum_{i=1}^{2^{|I|}} c_i (G_q(A_i^I) - R(A_i^I)) \right|$$

$$= \max_{I \subset \{1,\ldots,d\}, |I| = k} \quad \max_{c_1, \ldots, c_k \in \{-1,+1\}} \left| \sum_{i=1}^{k} c_i (G_q(A_i^I) - R(A_i^I)) \right|$$

For a given set of attributes I, the decision tree that maximizes $\left| \sum_{i=1}^k c_i (G_q(A_i^I) - R(A_i^I)) \right|$ is given by,

$$c_i = \begin{cases} +1, & G_q(A_i^I) \geq R(A_i^I) \\ -1, & G_q(A_i^I) \leq R(A_i^I), \end{cases} \tag{32}$$

which results in,

$$\sup_{c \in \mathcal{C}} |E_{G_q}[c(X_g)] - E_R[c(X_r)]| = \max_{I \subset \{1,\ldots,d\}, |I| = k} 2TV(\tilde{G}_{q_I}, \tilde{R}_I).$$

□

C. Implementation Details

This section provides comprehensive implementation details to ensure the reproducibility of our experimental results. We begin by describing our methodology for measuring MPR, including data preprocessing, classifier training, and attribute detection. We then detail the implementation specifics of baseline methods and our fine-tuning approach, including hyperparameter settings and optimization choices.

Details on MPR measurement. We utilized the FairFace training dataset to train classifiers for detecting gender, age, and race attributes. For gender classification, we maintained the original binary categories (male and female) from the dataset. Race classification preserved the seven original categories: White, Black, Southeast Asian, Middle Eastern, East Asian, Latino Hispanic, and Indian. For age classification, since classifying fine-grained age groups is challenging, we simplified the task by binarizing the labels into "young" (< 40 years) and "old" (≥ 40 years) categories to improve classification reliability. The classifiers were trained in the CLIP embedding space using linear classifiers, and the optimal models were selected through a grid search algorithm. As a result, the classifiers achieved accuracies of 98%, 95%, and 77% on the FairFace test dataset for gender, age, and race, respectively.

By default, we generated 1,000 images to measure MPR unless stated otherwise. We employed the Dlib face detector [34] to detect faces in the generated images and filtered to include only those images containing at least one detectable face. To estimate demographic attributes (gender, age, and race), we cropped the detected faces from the images. We applied our classifiers only to these cropped regions, which helps reduce noise from background elements. If multiple faces were detected in an image, we selected the largest face for analysis to ensure consistent evaluation. In contrast, for attribute-specific detection tasks (e.g., detecting the presence of a wheelchair or analyzing scene composition), we utilized the entire generated image without cropping to preserve all relevant visual information.

Implementation details of baselines and our finetuning method. All baseline methods were implemented using their publicly released codebases to ensure fair comparison. For each baseline, we maintained their default hyperparameter configurations as specified in their respective papers and repositories. To ensure consistent evaluation criteria, we adapted all baseline methods to utilize FairFace statistics rather than assuming equal representation when applicable.

As described in Section 5, our fine-tuning approach incorporates the MPR term (Equation 2) as a regularizer and leverages cached generations and functions c to enable efficient gradient computation (detailed in Algorithm 1). The hyperparameter configuration is as follows: we used a learning rate of 0.0005 with a warm-up phase and no subsequent scheduling, a mini-batch size of 8 per iteration, and applied fine-tuning exclusively to the text encoder of the SD v1.4 model using LoRA for $10,000$ iterations. The regularization strength λ was selected through a systematic search over $[0.1, 0.5, 1, 5, 10, 50, 100]$. To efficiently determine this value, we conducted preliminary training runs of 100 iterations each and selected the largest value that demonstrated consistent MPR improvements compared to the pre-fine-tuned model. In our case, $\lambda = 0.5$ was chosen. The buffer sizes for storing generated images B_{MPR} and functions B_C were both set to 32 to balance memory constraints with optimization stability.

Algorithm 1: Finetuning algorithm for achieving MPR

Input : Model θ, iteration I, mini-batch size B, MPR batch size B_{MPR}, $\widehat{C}$ size B_C, curation dataset $\mathcal{D}$
Set θ_0 from a pre-trained diffusion model
Set $\widehat{C}$ & $\widehat{P} \leftarrow [\,], [\,]$
for $t = 0$ **to** $T - 1$ **do**
  // Generate samples ;
  Generate B images X_B^t, X_B^o from θ_t, θ_0;
  $\widehat{P}$.extend(X_B^t) ;
  if $|\widehat{P}| > B_{MPR}$ **then**
    Pop some old images in $\widehat{P}$;
  // Calculate c ;
  $c_t \leftarrow \left| \frac{1}{k} \sum_{x_i \in \widehat{P}} c(x_i) - \frac{1}{m} \sum_{x_j \in \mathcal{D}} c(x_j) \right|$;
  $\widehat{C}$.append(c_t);
  if $|\widehat{C}| > B_C$ **then**
    Pop the oldest c in $\widehat{C}$;
  // Gradient update c ;
  $\theta_{t+1} \leftarrow \theta_t - \alpha \nabla \mathcal{L}_{\text{obj ver1}}(\theta_t)$;
Output : θ_T

D. Additional Experimental Results

In this section, we present detailed experimental analyses that complement and expand upon the results shown in our main paper. We examine the performance characteristics of our MPR framework across multiple dimensions, including (i.) comprehensive

Table 5. **MPR results for several traits**. Every MPR is obtained with groups of "Male", "Old" 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 by bootstrapping.

		Attractive	Emotional	Exotic	Poor	Terrorist	Thug
LCM-SDXL	DT($d=1$)	0.60 ($\pm$0.01)	0.44 ($\pm$0.01)	0.51 ($\pm$0.02)	0.44 ($\pm$0.01)	0.44 ($\pm$0.01)	0.58 ($\pm$0.02)
	DT($d=3$)	0.72 ($\pm$0.01)	0.63 ($\pm$0.01)	0.64 ($\pm$0.01)	0.66 ($\pm$0.01)	0.62 ($\pm$0.01)	0.69 ($\pm$0.01)
Stable Cascade	DT($d=1$)	0.57 ($\pm$0.01)	0.55 ($\pm$0.01)	0.46 ($\pm$0.01)	0.54 ($\pm$0.01)	0.83 ($\pm$0.00)	0.41($\pm$0.01)
	DT($d=3$)	0.81 ($\pm$0.00)	0.63 ($\pm$0.01)	0.64 ($\pm$0.01)	0.59 ($\pm$0.01)	0.87 ($\pm$0.00)	0.60 ($\pm$0.01)
Playground v2.5	DT($d=1$)	0.49 ($\pm$0.01)	0.41 ($\pm$0.01)	0.79 ($\pm$0.01)	0.74 ($\pm$0.01)	0.47 ($\pm$0.00)	0.66 ($\pm$0.01)
	DT($d=3$)	0.73 ($\pm$0.00)	0.55 ($\pm$0.01)	0.84 ($\pm$0.01)	0.80 ($\pm$0.01)	0.65 ($\pm$0.01)	0.85 ($\pm$0.00)
PixArt-Σ	DT($d=1$)	0.58 ($\pm$0.01)	0.55 ($\pm$0.01)	0.46 ($\pm$0.01)	0.54 ($\pm$0.01)	0.83 ($\pm$0.00)	0.41 ($\pm$0.01)
	DT($d=3$)	0.71 ($\pm$0.01)	0.63 ($\pm$0.01)	0.64 ($\pm$0.01)	0.59 ($\pm$0.01)	0.87 ($\pm$0.00)	0.60 ($\pm$0.01)
Splits of DT($d=3$) LCM-SDXL		Male Latino **White**	Old **White** Latino	Old Latino **Indian**	**Old** Black Indian	**Male** Old East Asian	Male Latino **Black**
Splits of DT($d=3$) Stable Cascade		Male Latino **White**	Old Latino **White**	**Male** Indian Latino	**Old** Middle Eastern Southeast Asian	Male Old **Indian**	**Male** Old East Asian
Splits of DT($d=3$) Playground v2.5		Old Latino **White**	**Male** Middle Eastern East Asian	Male Old **Indian**	Male **Indian** -	**Male** Indian White	Male Indian **Black**
Splits of DT($d=3$) PixArt-Σ		Male Old **White**	Male Latino **White**	Old **Indian** Black	**Old** Indian Black	**Male** Indian Middle Eastern	Male **White** Black

trait representation analysis using additional state-of-the-art diffusion models (ii.) analysis of empirical vs. true MPR gaps across different function classes, (iii.) quantitative examination of contextual representation in historical settings, such as the "computer programmer for ENIAC" and (iv.) supplementary qualitative results. Each subsection provides experimental evidence supporting our theoretical framework while revealing new insights about representational biases in text-to-image systems.

D.1. MPRs for Traits on Advanced Text-to-Image Models

While our main paper examined trait biases across three baseline models (SD v1.4, SD v2.1, SDXL), here we extend this analysis to include four state-of-the-art models: LCM-SDXL [41], Stable Cascade [52], Playground v2.5 [36], and PixArt-Σ [14]. This broader analysis helps understand whether other recent architectural and training advances have addressed representational biases.

Table 5 presents MPR measurements using identical experimental settings as Table 1 in the main paper. Our findings reveal several key insights:

- Despite their documented improvements in image quality and generation capabilities, these newer models exhibit bias levels comparable to or sometimes exceeding those of SDXL across all six examined traits.
- We observe consistent bias patterns across all models regardless of their architectural differences. For example, decision trees consistently select "white" as a splitting criterion when evaluating "attractive" trait bias and "black" when evaluating "thug" trait bias. This consistency suggests these biases may stem from deeper societal stereotypes present in training data rather than specific architectural choices.

These findings emphasize that advances in model architecture and image generation quality do not automatically translate to improved representational fairness. They underscore the critical need for explicitly incorporating intersectional fairness considerations into model development rather than treating them as properties that will naturally emerge from technical improvements.

D.2. Maximum Gap between Empirical and True MPRs Depending on the Function Class

In Proposition 1, we established the relationship between the number of images of $\widehat{G}_q$ and $\widehat{R}$, the complexity of the function class $\mathcal{C}$, and the approximation error of MPR$(\mathcal{C}, \widehat{G}_q, \widehat{R})$. Specifically, we demonstrated that, in the worst case, there is a trade-off between function complexity and the number of images to achieve the same level of accuracy in MPR estimation.

To support this experimentally, we analyzed how the accuracy of MPR estimation changes with respect to function

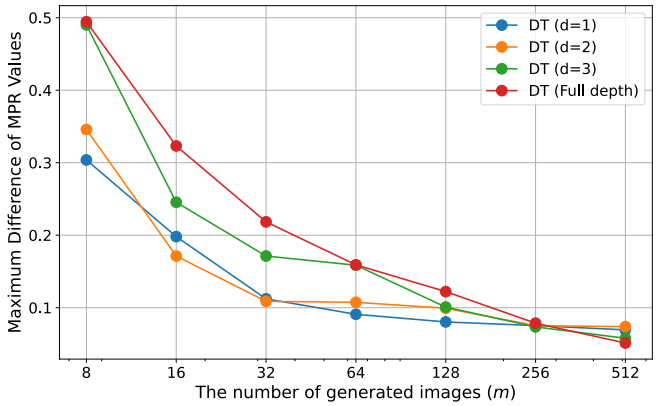

Figure 4. **Maximum deviation between the estimated MPRs and the true MPR depending on decision tree depth and the number of generated images.** The number following "DT" indicates the depth of the decision tree, and the maximum deviation is calculated after 30 repetitions of the estimation process.

Table 6. **MPR values of different diffusion models on a balanced (left) and contextual (right) curation set.**

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)
Entigen [9]	0.04 ($\pm$ 0.02)	0.33 ($\pm$ 0.01)
UCE [24]	0.19 ($\pm$ 0.01)	0.48 ($\pm$ 0.01)
ITI-GEN [66]	0.33 ($\pm$ 0.01)	0.83 ($\pm$ 0.01)

complexity and the number of images when estimating the MPR value for the CEO using the SD v1.4 model. To calculate the approximation error for each trial of estimating an MPR value, *i.e.*, the LHS in eq. 7, we assume that the MPR value calculated using 5,000 generated images is an accurate approximation of the true MPR. For the given number of generated images $-m-$ we generated images, estimated the MPR, and repeated this process 30 times. We then reported the maximum deviation from the assumed true MPR. In this experiment, we used the FairFace test dataset as the reference data and calculated MPR using gender, age, and race group labels with decision tree functions.

As shown in Figure 4, the deviation from the true MPR consistently decreases as the number of images increases. Furthermore, we observe that the deviation becomes larger when using deeper decision trees, *i.e.*, more complex functions. These findings directly support the arguments in Proposition 1, implying that while MPR can capture more nuanced biases with higher function complexity, the cost of obtaining sufficient images to maintain accuracy also increases.

D.3. MPR Results of Other Baseline Methods for "Computer programmer for ENIAC"

Building upon the results presented in Table 3 of the main paper, we evaluated the performance of additional baseline methods in representing ENIAC programmers. Table 6 presents MPR values measured against both equal gender distribution and historically accurate (Google Images) reference sets. Our expanded analysis reveals:

- All baseline methods successfully reduce bias when evaluated against equal gender representation, with FairDiffusion and Entigen achieving particularly low MPR scores (0.03 and 0.04, respectively).
- However, when evaluated against the historically accurate reference distribution (which consists solely of female programmers), all methods show substantial representational disparities. Even the best-performing methods (FairDiffusion and Entigen) exhibit MPR values above 0.30. One reason for the increase in MPR with non-uniform reference statistics in existing diffusion models (even with fairness interventions) is that they are trained to ensure equal representation and not contextual representation.
- Notably, ITI-GEN, which performs well in other contexts, shows the highest disparity (0.83) against historical accuracy, suggesting that methods optimized for general fairness might inadvertently work against accurate historical representation.

These results highlight a fundamental tension between different fairness objectives: while these methods effectively promote

Figure 5. Qualitative comparison of images generated for the prompt *a portrait photo of a {pilot, chef, flight attendant, housekeeper, taxi driver, nurse, therapist}* across different methods. (Top row) Images generated by the original SD v1.4 model show a strong bias toward white male pilots, lacking demographic diversity. (Middle row) Images generated by FairDiffusion achieve better gender balance but still show limited representation across intersectional groups. (Bottom row) Images generated by SD v1.4 fine-tuned with MPR.

gender equality in general, they may inadvertently diminish historically significant representations of women in computing. This underscores the importance of context-aware evaluation metrics and the need for more nuanced approaches to bias mitigation.

D.4. Qualitative Results

In this section, we show that an MPR-optimized model demonstrates significant improvements in achieving balanced representation across intersectional groups. Specifically, we fine-tuned Stable Diffusion v1.4 using our MPR-based optimization approach to evaluate this capability. As shown in Figures 5 and, when generating images for the prompt "a portrait photo of a {pilot, chef, flight attendant, housekeeper, taxi driver, nurse, therapist}" our fine-tuned model produces diverse, high-quality images that span multiple demographic dimensions, including gender, age, and race. While the original SD v1.4 predominantly generates images of white male pilots, and FairDiffusion primarily addresses gender balance, our MPR-optimized version effectively captures broader intersectional diversity. The qualitative results in Figure 5 and 6 (split into two figures for formatting purposes) validate our quantitative findings, showing that MPR optimization can successfully balance representation across demographic groups while preserving image quality and maintaining appropriate professional context.

D.5. Sensitivity Analysis on the Regularization Strength in Our Finetuning Method

We investigate how the regularization strength λ in our objective affects the performance of the model, in terms of representational fairness and image quality. Table 7 presents the MPR values and CLIP scores for different values of $lambda$. As λ increases, we observe a consistent decrease in MPR and an increase in CLIP. This trend demonstrates that the regularization term effectively guides the model toward generating more representative and semantically coherent outputs.

E. Discussion on the practical selection of sample sizes in the MPR framework

We identify two primary real-world scenarios where the MPR framework can be utilized to evaluate text-to-image models. In the first scenario, we assess whether a generative model f sampling from distribution G meets a specified fairness threshold

Figure 6. Continuation of Figure 5. Qualitative comparison of images generated for the prompt *a portrait photo of a {housekeeper, taxi driver, nurse, therapist}* across different methods.

Table 7. MPR and CLIP scores by varying λ of our finetuning method

λ	$\left(\frac{1}{2}\right)^{\frac{3}{2}}$	$\left(\frac{1}{2}\right)^{1}$	$\left(\frac{1}{2}\right)^{\frac{1}{2}}$	$\left(\frac{1}{2}\right)^{0}$	SD v1.4
MPR (GAR) (↓)	0.235	0.252	0.267	0.335	1.89
CLIP score (↑)	0.292	0.300	0.303	0.306	0.311

ρ (e.g., a representation target mandated by policymakers) by testing if $\mathrm{MPR}(\mathcal{C}, G, R) \leq \rho$. In the second scenario, we compare two different generative models f and g, sampling from distributions G and G' respectively, by evaluating whether $\mathrm{MPR}(\mathcal{C}, G, R) \leq \mathrm{MPR}(\mathcal{C}, G', R)$ and vice versa. For the threshold comparison scenario, we employ a one-sided t-test to determine if the model achieves representation below ρ. For model comparison, we utilize a two-sided t-test to assess relative performance between models. In both cases, when the estimated MPR value approaches ρ or when two models yield similar MPR values, higher precision estimates become crucial, necessitating lower standard deviation in our measurements.

As established in Proposition 1, we can achieve this increased precision by expanding the number of sampled images. This allows MPR users to control the trade-off between computational cost and estimation accuracy by adjusting the number of generated and reference images based on their specific requirements. However, since the true distributions G and R are unknown in practice, we cannot directly calculate the standard deviation. Instead, we employ bootstrapping to estimate these values empirically. Figure 7 visualizes the relationship between sample sizes and estimation precision, showing standard deviation values estimated through bootstrapping as a function of the number of generated –k– and reference –m– images.

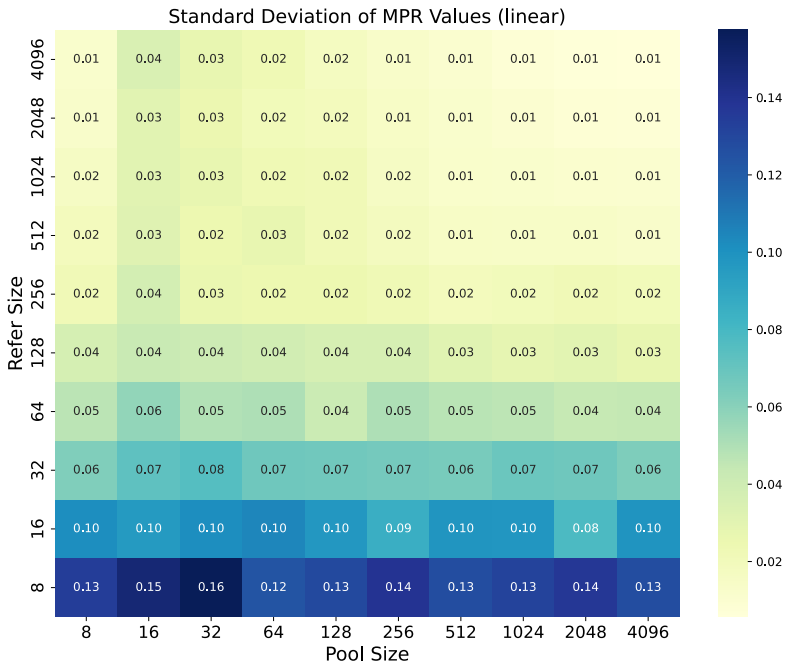

Figure 7. **Heatmap of MPR standard deviations obtained by bootstrapping.** Each value represents the MPR calculated using a linear classifier function class, considering gender, age, and race as group attributes.

The observed decrease in standard deviation with increasing k and m aligns with our theoretical predictions from Proposition 1, providing practical guidance for sample size selection in MPR evaluation.

Figure 8. **Images generated from the "CEO" prompt using the vanilla SD v1.4 model.** The leftmost and second color bars for each image represent the estimated gender and race, respectively. Gender is indicated by blue for males and red for females. Race is represented by the following colors: grey for White, black for Black, yellow for East Asian, green for Southeast Asian, orange for Indian, brown for Latino/Hispanic, and purple for Middle Eastern.

Figure 9. **Images generated from the "CEO" prompts using FairDiffusion.** The color details follow the same scheme as Fig. 8.

Figure 10. **Images generated from the "disability" prompts using the vanilla SD v1.4.** The leftmost and second color bars for each image represent the estimated presence of a wheelchair and race, respectively. The presence of a wheelchair is indicated by blue for wheelchair users and red for non-wheelchair users. Race is represented by the following colors: grey for White, black for Black, yellow for Latino/Hispanic, and purple for Asian.