Fair Text-to-Image Diffusion via Fair Mapping

Jia Li[1, 2, 4*], **Lijie Hu**[1, 2, 3*], **Jingfeng Zhang**[2, 5], **Tianhang Zheng**[6,7], **Hua Zhang**[4], **Di Wang**[1, 2, 3 †]

[1]Provable Responsible AI and Data Analytics (PRADA) Lab
[2]King Abdullah University of Science and Technology
[3]SDAIA-KAUST
[4]Institute of Information Engineering, Chinese Academy of Sciences
[5]University of Auckland
[6]The State Key Laboratory of Blockchain and Data Security, Zhejiang University
[7]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
di.wang@kaust.edu.sa

Abstract

In this paper, we address the limitations of existing text-to-image diffusion models in generating demographically fair results when given human-related descriptions. These models often struggle to disentangle the target language context from sociocultural biases, resulting in biased image generation. To overcome this challenge, we propose Fair Mapping, a flexible, model-agnostic, and lightweight approach that modifies a pre-trained text-to-image diffusion model by controlling the prompt to achieve fair image generation. One key advantage of our approach is its high efficiency. It only requires updating an additional linear network with few parameters at a low computational cost. By developing a linear network that maps conditioning embeddings into a debiased space, we enable the generation of relatively balanced demographic results based on the specified text condition. With comprehensive experiments on face image generation, we show that our method significantly improves image generation fairness with almost the same image quality compared to conventional diffusion models when prompted with descriptions related to humans. By effectively addressing the issue of implicit language bias, our method produces more fair and diverse image outputs.

Introduction

Text-to-image diffusion models [Ho, Jain, and Abbeel 2020, Song, Meng, and Ermon 2021] have achieved remarkable performance in various applications [Dhariwal and Nichol 2021, Wu et al. 2022, Ceylan, Huang, and Mitra 2023, Blattmann et al. 2023, Poole et al. 2023, Tevet et al. 2023] by involving incorporation of textual conditional elements with language models [Ho and Salimans 2022]. With the increasing popularity among the public, the generation of diverse images [Ouyang, Xie, and Cheng 2022] about human-related description becomes crucial, yet it remains a challenging task [Ning, Li, and Jianlin Su 2023, Schramowski et al. 2023, Bansal et al. 2022]. For example, when prompted to "An image of a computer programmer" and "An image of a confident person" to Stable Diffusion, as exemplified in Figure 1, it is obvious that the majority of results are white male, even if there are no specified gender or race descriptions in textual condition.

Figure 1: Fair Mapping balances demographic biases in text-to-image models by minimally adjusting training to generate more diverse images compared to outputs of Stable Diffusion.

This phenomenon demonstrates that text-to-image diffusion models learned implicit correlations between demographic visual representations and textual descriptions during training. Real-world data is often biased and incomplete, reflecting inherent stereotypes in human perceptions [Makady et al. 2017, Bansal et al. 2022]. When training on a dataset that includes pairs of images and texts [Schuhmann, Beaumont, and Richard Vencu 2022, Schuhmann, Vencu, and Romain Beaumont 2021], text-to-image diffusion models face challenges in disentangling sensitive attributes such as gender or race, from specific prompts, which potentially introduce during denoising and lead to societal biases in generation.

Consequently, the pressing question remains: *How can unbiased inferences be made in the face of training data that is inherently biased?* Nowadays, [Orgad, Kawar, and Belinkov 2023, Gandikota, Orgad, and Belinkov 2024] highlight text-to-image models implicitly assume sensitive attributes from the language model, where textual condition embeddings demonstrate an inclination towards specific demographic groups. As a result, some work [Shen, Du, and Pang 2023, Fraser, Kiritchenko, and Nejadgholi 2023] takes fine-tune methods to control the distribution of generated images. However, updating the parameters of the diffusion model with expensive data collection potentially results in

*These authors contributed equally.
†Corresponding author

knowledge forgetting, ultimately detrimental to stable generative ability. [Friedrich, Schramowski, and Manuel Brack 2023] allows human instruction, enabling to guide output based on desired criteria. These post-processing methods require significant computational resources and time, which is not efficient enough without a robust framework.

In this paper, we propose a novel post-processing, model-agnostic, and lightweight method namely **Fair Mapping**. Briefly speaking, there are two additional components in Fair Mapping compared to vanilla diffusion models: The first one is a linear mapping network which is strategically designed to rectify the implicit bias in the representation space of text encoder in text-to-image diffusion models. It addresses the disentanglement of the target language context from implicit language biases by introducing a fair penalty mechanism. This mechanism fosters a harmonious representation of sensitive information within word embeddings via a flexible linear network with only a modest addition of new parameters. At the inference stage, Fair Mapping introduces a detector, which aims to detect input containing potential biased content for a robust generation. To summarize, our contributions are three-fold.

- *Analysis of Language Bias and Proposed Fairness Evaluation Metric:* We first quantitatively explain the bias issue in generated images caused by the textual condition within text-guided diffusion models, providing insights into the contributing dynamics. We introduce evaluation metrics designed to assess the language bias and diffusion bias of diffusion models in generating text-guided images. These metrics provide a systematic and objective measure for quantifying the reduction of bias in the generative process, enabling a more precise evaluation of fairness outcomes.
- *Innovative Fair Mapping Module:* To mitigate language bias, we develop a model-agnostic and lightweight method namely Fair Mapping. Generally speaking, Fair Mapping introduces a linear network before the Unet structure, which optimizes minimal extra parameters for training, enabling seamless integration into any text-to-image generative models while keeping their parameters unchanged. At the inference stage, there is an additional detector compared to conventional diffusion models, which aims to detect whether input text prompt given by users pass through the linear network for debias.
- *Comprehensive Experimental Evaluations:* Finally, we conduct comprehensive experiments of our methods to ensure our generated images' fairness and quality. Specifically, our experiments demonstrate that our method improves performance and outperforms several text-to-image diffusion models based on human descriptions for fairness, while the image quality is very close to and even better than that of the base diffusion models.

Related Work

Text-guided Diffusion Models. Text-guided diffusion models merge textual descriptions with visual content to create high-resolution, realistic images that align with the semantic guidance provided by the accompanying text prompts [Ramesh, Dhariwal, and Alex Nichol 2022, Saharia, Chan, and Saurabh Saxena 2022, El-Nouby et al. 2018, Kim et al. 2023, Avrahami et al. 2023, Balaji, Nah, and Xun Huang 2022, Feng et al. 2023, He, Salakhutdinov, and Kolter 2023]. However, this fusion of modalities also brings to the forefront issues related to bias and fairness [Struppek, Hintersdorf, and Kersting 2022, Bansal et al. 2022], which have prompted extensive research efforts to ensure that the generated outputs do not perpetuate societal inequalities or reinforce existing biases. In this paper, we explore these challenges and the state-of-the-art solutions aimed at enhancing the fairness and equity of text-guided diffusion models.

Bias in Diffusion Models. While large datasets are commonly used in data-driven generative models, they often contain social and cultural biases [Birhane, Prabhu, and Kahembwe 2021, Schuhmann, Beaumont, and Richard Vencu 2022, Schuhmann, Vencu, and Romain Beaumont 2021]. Previous efforts have addressed this challenge by optimizing model parameters after training [Shen, Du, and Pang 2023, Fraser, Kiritchenko, and Nejadgholi 2023], while [Jiang, Lyu, and Ma 2023] accomplish distributional control by updating the latent code. [Friedrich, Schramowski, and Manuel Brack 2023] introduces a post-processing mechanism based on human instructions. [Chuang, Jampani, and Li 2023] remove biased directions in text embeddings to mitigate bias in vision-language. Similarly, [Orgad, Kawar, and Belinkov 2023, Gandikota, Orgad, and Belinkov 2024] update the model's cross-attention layers to achieve concept editing in certain text for debiasing. In our work, we also seek to bridge this gap by addressing language biases in semantic representation space, thereby contributing to a more comprehensive understanding and mitigation of biases in generative data by employing an end-to-end framework.

Bias in Language Models. The Transformer structure of language models is capable of storing and encoding knowledge, including societal biases reflected in the training data. This capability is also extended to text-to-image diffusion models through the incorporation of attention layers [Meng, Sharma, and Andonian 2022, Arad, Orgad, and Belinkov 2023, Berg, Hall, and Yash Bhalgat 2022]. Ensuring fairness in these models has been extensively studied, especially in the context of large-scale models [Gehman et al. 2020, Abid, Farooqi, and Zou 2021, Bender et al. 2021, Zhang, Wang, and Sang 2022, Radford, Kim, and Chris Hallacy 2021, Tian, Lai, and Moore 2018, Ding et al. 2021]. Efforts have been made to address these biases with approaches [Wang, Zhang, and Sang 2022, Dehouche 2021] aiming to mitigate the impact of bias. In our work, we explore the intersection of bias mitigation efforts in language models, which is a critical juncture in the pursuit of fairness and ethics in artificial intelligence.

Language Bias in Text-to-Image Diffusion Models

For starters, we list key notations that will be used throughout the paper.

Notations. Consider a keyword set C such as a set of different occupations. For each keyword c, such as "doctor", it has a set A of possible sensitive attributes such as "male" or "female". Note that here we suppose the attribute set is

the same for all keywords for convenience. For language bias, we denote $prompt(a, c)$ for each $a \in A, c \in C$ as a prompt in a uniform and specific format. For example, $prompt(\text{male}, \text{doctor})$ = "an image of a male doctor". We also denote $prompt(\text{`` '}, c)$ as the prompt where there is no sensitive attribute, such as $prompt(\text{`` '}, \text{doctor})$ = "an image of a doctor". Given the text encoder for textual conditioning in a text-to-image diffusion model, we extract text representations f and $\{f_j\}_{j=1}^{|A|}$ from $prompt(\text{`` '}, c)$ and $\{prompt(a_j, c)\}_{a_j \in A}$ respectively. These representations are essential in conventional text-guided diffusion models for generating coherent and contextually relevant text samples.

Before showing our method for mitigating language bias, it is necessary to address the following fundamental question: *Whether there indeed exists implicit language bias even if there is no explicit sensitive attribute in the textual information of prompt?* Based on the above notation, we propose the bias metric for the input prompts and the generated outputs of text-to-image diffusion models.

1) **Diffusion Bias.** We propose a novel evaluation metric based on group fairness to robustly assess fairness in generative results of diffusion models across diverse groups. This metric captures variations in generated outcomes among demographic attributes [Hardt, Price, and Srebro 2016], such as gender, and quantifies fairness by evaluating equilibrium. Our study adopts a highly specific and constrained definition of fairness in the evaluation process. A diffusion model is absolutely fair if it satisfies that for any keyword $c_k \in C$

$$P(s = a_i | c = c_k) = P(s = a_j | c = c_k), \text{for all } a_i, a_j \in A, \quad (1)$$

where s is the random variable of the sensitive attribute for the output of the diffusion model, c is the random variable of the keyword in conditional textual information, $P(s = a_i | c = c_k)$ represents the probability of the sensitive attribute s of generative images expressing a_i given a specific conditional prompt with keyword c_k. We define our diffusion bias evaluation criteria towards attribute a_i for c_k as follows:

$$DBias_{a_i}(c_k) = P(s = a_i \mid c = c_k) - \frac{1}{|A|} \sum_{a_j \in A} P(s = a_j \mid c = c_k) \quad (2)$$

Thus, based on (1), for a keyword c_k, our metric on the diffusion bias is defined as follows:

$$\text{BiasScore}(c_k) = \sqrt{\frac{1}{|A|} \sum_{a_i \in A} (DBias_{a_i}(c_k))^2}. \quad (3)$$

Thus, for a dataset C that contains keywords, our fair evaluation metric on the diffusion bias is $\frac{1}{|C|} \sum_{c_k \in C} \text{BiasScore}(c_k)$. A smaller value indicates that the method is fairer.

2) **Language Bias.** We assess language bias by incorporating semantic similarity calculation [Chen, Kornblith, and Mohammad Norouzi 2020, Mikolov, Chen, and Greg Corrado 2013] between $prompt(a, c)$ and $prompt(\text{''}, c)$. Specifically, we use Euclidean distance to evaluate the distance between prompt with and without explicit sensitive attributes. The closer distance indicates a potential bias towards one specific sensitive attribute. We define our language bias evaluation

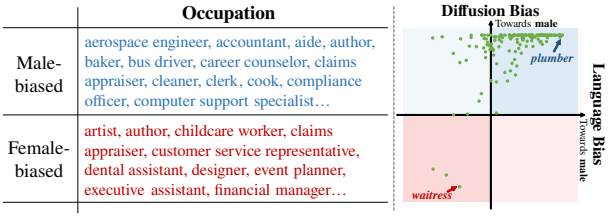

Figure 2: **Language Bias and Diffusion Bias Visualization.** We conduct a bias analysis of language characteristics and generated outcomes during the diffusion process. Left: Examples of language prejudice. Right: Language bias and diffusion bias for occupational data. Each point represents an occupation.

criteria towards attribute a_i for keyword c_k and our input prompts as $LBias_{a_i}(c_k)$:

$$LBias_{a_i}(c_k) = -\|f_j - f\|_2 + \frac{1}{|A|} \sum_{a_j \in A} \|f_j - f\|_2, \quad (4)$$

where $\|f_j - f\|_2$ represents the Euclidean distance between the prompt generated with the sensitive term a_i and the keyword c_k, compared to the prompt generated with no sensitive term. Thus, the total language bias for keyword c_k and our prompt is $\frac{1}{|A|} \sum_{a_i \in A} LBias_{a_i}(c_k)$.

Based on the above two metrics, we conduct following experiments in the occupation keyword set listed in Appendix: 1) First, we calculate language bias on every occupation over sensitive attribute gender. 2) Then, for each occupation c, we use the following prompt format $prompt(\text{`` '}, c)$ for guiding the stable diffusion model to generate 100 images and measure the diffusion bias.

Figure 2 shows the experimental results. In the left region of Figure 2, we analysed language bias on specific occupations and provided examples to illustrate some samples aligning with societal stereotypes. For example, our analysis of the term "aerospace engineer" highlights a clear gender bias favoring males, reflecting the gender stereotype associated with this profession in the real world. The right-hand side of Figure 2 displays the biases associated with various occupations by creating a scatter plot with diffusion bias on the y-axis and language bias on the x-axis. Moreover, we discovered the majority of data are concentrated in where a male bias was revealed in both diffusion and language bias. This research implies that language bias and diffusion bias are mutually reinforcing, which infuse in the cross-attention layers of the UNet. In summary, implicit language bias is one of the direct factors leading to diffusion bias. From the view of language assumption, it is reasonable to reduce the impact it has on diffusion bias and promote more equitable and unbiased generative outcomes.

Mitigating Implicit Bias via Fair Mapping

Motivated by above analysis, we introduce Fair Mapping to disentangle the implication of language from diffusion generation. Generally speaking, Fair Mapping introduces two additional components as a post-processing method on well-trained text-to-image diffusion models: A linear network named Fair Mapping and a detector that will be activated during the inference stage. Training and inference procedures of

Fair Mapping are elucidated in Figure 3. We implement Fair Mapping consisting of linear stacking networks, drawing inspiration from StyleGAN [Karras, Laine, and Aila 2018] and MixFairFace [Wang et al. 2022], which enables the correction of native language semantic features, ultimately leading to their debiasing and alignment with the balanced embedding space. The detector at the inference stage is used for robust generation to decide whether the input will be debiased. We provide details on the training and inference stages hereafter.

Training Fair Mapping Network

Recall that our linear network Fair Mapping, denoted as M, is to transform the representation of the input prompt to a debiased space. Our idea is to maintain the representation aligned with a balanced state of sensitive attributes, so they can introduce little assumption to visual representation. First, given a keyword dataset and group of sensitive attributes, we will construct two distinct types of prompts that have a consistent and uniform format for each keyword c. The first type constitutes the original input prompt, denoted as $prompt(\text{`` ''}, c)$, where we explicitly exclude any sensitive attributes (represented as `` ''). It does not prioritize explicit sensitive attribute information. In contrast, the second type of prompt, $prompt(a_j, c)$, where $a_j \in A$, incorporates specific sensitive words. These prompts are designed to quantitatively explore the language relationship between sensitive attributes and the target keyword. Moreover, we have language representation vectors f and $\{f_j\}_{j=1}^{|A|}$ from $prompt(\text{`` ''}, c)$ and $\{prompt(a_j, c)\}_{a_j \in A}$ respectively. Given these, our Fair Mapping is after the pre-trained text encoder and further transforms these representation vectors: $v = M(f), v_j = M(f_j)$ for $a_j \in A$.

Our aim with this transformation process is to ensure that the representations are fair and unbiased, and exhibit equitable treatment across sensitive attributes. In detail, the objectives of v and $\{v_j\}_j$ are two-fold: 1) They should maintain semantic consistency akin to f and $\{f_j\}_j$ respectively, serving as keeping the original contextual information. 2) More importantly, v should equalize the representation of different demographic groups and prevent the encoding of implicit societal biases. Therefore, we employ bias-aware objectives and regularization techniques to guide representations to be balanced from sensitive information. Below we will discuss how to achieve the above two goals.

Keeping semantic consistency. Our general idea is designed to maintain consistency and semantic coherence between the original embeddings and mapped embeddings. To achieve this objective, we employ a strategy for minimizing the disparity between pre-transformed and post-transformed features in the embedding space. Specifically, we adopt the mean squared error (MSE) as a metric to measure the reconstruction error, drawing inspiration from [Kingma and Welling 2014]. By applying this metric, we compute a semantic consistency loss for each keyword:

$$\mathcal{L}_{text} = \frac{1}{|A|+1}\left(||v-f||_2^2 + \sum_{a_j \in A} ||v_j - f_j||_2^2\right). \quad (5)$$

Through the minimization of this loss over all keywords, we can ensure that the mapped embeddings preserve the crucial information and semantic attributes inherent in the original embeddings. This process safeguards the fidelity and integrity of the data throughout the mapping transformation.

Fairness loss. Distances of embeddings to groups with sensitive attributes can inadvertently encode demographic information. For example, if the word "doctor" is closer to "male" than "female" in the representation space of language models, it may inherently convey gender bias [Chen, Kornblith, and Mohammad Norouzi 2020]. To mitigate this issue, we employ an invariant loss that entails the adjustment of associations between sensitive attributes and naive prompts during the training process using mapping offsets. The primary objective is to diminish associations of implicit sensitive attributes with text embeddings through the application of mapping offsets, promoting a more unbiased representation.

To equalize the representations of attributes, we ensure that representations of prompts have balanced distances from the representations of prompts with specific sensitive attributes, thereby reducing the bias assumptions in the semantic space. In the case where the size of the sensitive group A is 2, we can minimize the difference in distance between the native embeddings, expressed as $|d(v, v_1) - d(v, v_2)|$. Here, $d(\cdot, \cdot)$ represents the Euclidean distance [Dokmanic et al. 2015] between embeddings. To address the computational complexity when dealing with a large attribute set containing multiple sensitive groups, we can optimize representation bias by reducing the variance in the distance between embeddings instead of calculating the difference in the distance for each pair. The fairness loss term denoted as $\mathcal{L}_{fair}$ can be formulated as follows:

$$\mathcal{L}_{fair} = \sqrt{\frac{1}{|A|}\sum_{a_j \in A}\left(d(v, v_j) - \overline{d(v, \cdot)}\right)^2}. \quad (6)$$

Here, $d(v, v_i)$ represents the Euclidean distance between the native embedding v and the specific sensitive attribute embedding v_i. $\overline{d(v, \cdot)}$ refers to the average distance between the native embedding v and all the sensitive attribute embeddings v_j. By incorporating this fairness loss term into the training objective, we aim to minimize the variance in the distance between v and v_j with sensitive attributes. To optimize the overall objective, we combine the semantic consistency loss, denoted as $\mathcal{L}_{text}$ (from (5)), with the estimated bias difference from the fairness penalty (from (6)). This results in the following combined loss function for each keyword:

$$\mathcal{L} = \mathcal{L}_{text} + \lambda \mathcal{L}_{fair}, \quad (7)$$

where λ is a hyperparameter that controls the trade-off between semantic consistency and fairness. By minimizing this combined loss function, we aim to simultaneously ensure semantic consistency and reduce bias in the language representation. We optimize the parameters of Fair Mapping while keeping the parameters of the diffusion model fixed.

Inference

In the inference stage, as Figure 3 demonstrated, the detector is introduced before Fair Mapping. To build a robust system for users, the main goal of the detector is to decide whether the text should pass or skip Fair Mapping. In detail, it aims

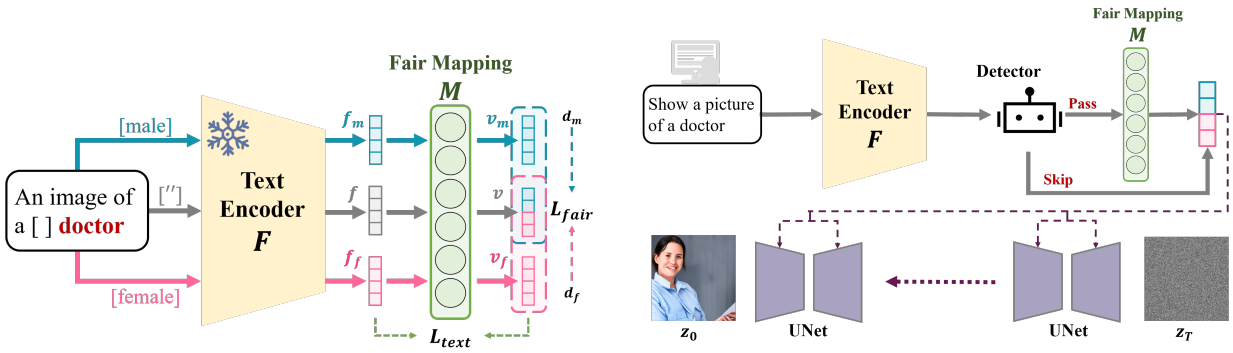

Figure 3: **Left:** In the training stage, the parameters of the text encoder are frozen, and we apply $\mathcal{L}_{text}$ and $\mathcal{L}_{fair}$ to update Fair Mapping. d_a denotes the distance between v_a and v. **Right:** In the inference stage, the detector after the text encoder determines whether the text should pass or skip the Fair Mapping linear network.

to match the target input with implicit sensitive attributes and avoid solving with explicit sensitive attributes. Due to the space limit, details for the detector can be found in Appendix.

Discussions. We can easily see that there are several strengths of Fair Mapping. Firstly, Fair Mapping is model-agnostic, i.e., as it only introduces a linear mapping network and a detector after the text encoder, it can easily be integrated into any text-to-image diffusion model as well as be a plug-in text-encoder with the same parameters. Secondly, Fair Mapping is lightweight. As a post-processing approach, Fair Mapping only introduces an additional linear map to be optimized while keeping the parameters of the diffusion model fixed. Moreover, as we will mention in the experiments, an eight-layer linear network is sufficient to achieve good performance on both utility and fairness with little additional time cost. Finally, our method is quite flexible. Due to the simplicity of our loss for each keyword, our linear network can be customized for any other prompts, loss of semantic consistency, and loss of fairness.

Experiments

We mainly compared our Fair Mapping (FairM) with three text-guided diffusion models: Stable Diffusion (SD) [Rombach, Blattmann, and Dominik Lorenz 2022], Structured Diffusion (StruD) [Feng, He, and Tsu-Jui 2023] and Composable Diffusion (ComD) [Tang, Yang, and Chenguang 2023], as well as three state-of-the-art methods for fair diffusion generation: Fair Diffusion (FairD) [Friedrich, Schramowski, and Manuel Brack 2023], Unified Concept Editing (UCE) [Gandikota, Orgad, and Belinkov 2024] and Debias Visual Language (debiasVL) [Chuang, Jampani, and Li 2023]. We report the performance of our models from three aspects: Our Fair Mapping 1) outperforms baselines in fairness evaluation and computation cost, 2) showcases alignment and diversity in human-related descriptions by quantitative analysis, 3) is more lightweight and introduces acceptable time overhead 4) matches human preferences in image quality and text alignment of state-of-the-art text-to-image diffusion methods.

Experimental Setup

Datasets. We select a total of 150 occupations and 20 emotions for the fair human face image generation following [Ning, Li, and Jianlin Su 2023]. For sensitive groups, we choose gender groups (male and female), racial groups (Black, Asian, White, Indian) and Age (young, middle age, old) provided by [Kärkkäinen and Joo 2021]. We provide a comprehensive list of keywords in Appendix.

Implementation details. In our experiments, we use stable Diffusion v1.5 [Rombach, Blattmann, and Dominik Lorenz 2022] as the base model and implement 50 DDIM denoising steps for generation. Standardized prompts are used for occupations and emotions. More details are in Appendix.

Evaluation metrics. Besides the two evaluation metrics in the above Section , language bias and diffusion bias (we denote as Bias), below we introduce some metrics for the utility and the quality of the generated images.

1) **Alignment:** To measure the alignment between generated images and human-related content, we adopt the CLIP-Score [Hessel, Holtzman, and Maxwell Forbes 2021], which measures the distance between input textual features and generated image features. Due to the limitation [Otani, Togashi, and Yu Sawai 2023] of capturing specific requirements of human-related textual generation, we introduce the Human-CLIP metric (see Appendix for details) focusing on evaluating the CLIP-Score related to human appearance.

2) **Diversity:** We use intra-class average distance (ICAD) [Le and Odobez 2018] to evaluate the diversity of generative results. For each keyword, we measure the average distance between all generated images and the center of images by using a distance metric by squared Euclidean distance (see Appendix for details). A lower intra-class average distance suggests that the generative results are more similar and excessively focus on a few specific samples.

Experimental Results

Figure 4 demonstrates the effectiveness of our method for debiasing. When compared to existing text-guided diffusion approaches, Fair Mapping demonstrates an enhanced capacity for generating diverse sensitive groups while upholding the stability of the generated results.

Fair Mapping can mitigate bias in diffusion models. The comparative analysis depicted in Figure 1 in our Appendix underscores the significant strides made by our method in mitigating language bias and diffusion bias. Besides a general

Figure 4: Comparison with original SD and different debiasing methods in prompt "an image of an engineer". Our method makes generated images equally represent genders and races. More visual results are in Appendix.

illustration of language bias and diffusion bias. Table 1 showcases the fair evaluation results for sensitive attributes: gender and race. Fair Mapping demonstrates significant improvements in fairness compared to other baseline approaches. It achieves lower Diffusion Bias compared to Stable Diffusion for all sensitive attribute groups, especially on gender (18% for Occupation and 54% for Emotion) and race (14% for Occupation and 38% for Emotion). Besides, compared with other debiasing methods, Our method outperforms in enhancing the representation of minority groups. We did not test for Fair Diffusion primarily because the generated results heavily depend on subjective post-processing by humans.

Dataset	Models	Bias (G.)	Bias (R.)	Bias (A.)
Occupation	SD	0.4466	0.4652	0.3173
	StruD	0.4141	0.4100	0.2911
	ComD	0.4027	0.4203	0.2884
	UCE	0.3802	0.3266	0.2688
	debiasVL	0.4573	0.4178	0.2982
	FairM(Ours)	**0.3625**	**0.2113**	**0.2547**
Emotion	SD	0.2599	0.1893	0.2142
	StruD	0.2368	0.1824	0.2529
	ComD	0.2344	0.1489	0.2318
	UCE	0.2314	0.1505	0.1993
	debiasVL	0.2992	0.1592	0.2231
	FairM(Ours)	**0.2231**	**0.1178**	**0.1561**

Table 1: Fair evaluation results of sensitive group gender and race. G., R., and A. denote Gender, Race and Age, respectively. The **bold** value denotes the best performance.

Alignment and diversity. In Table 2, when comparing with baselines for text-to-image diffusion generation, debiasing methods restrict the model's ability to generate images as a trade-off for fair generation. Although both UCE methods maintain the quality of generation, they perform poorly in Human-CLIP evaluation. Furthermore, our method showcases a strong alignment with human-related prompts, with 13% improvements in the Human-CLIP metric for occupation. our method successfully captures human-related characteristics in generated images, despite having a slight loss in CLIP-Score.

Models	Occupation			Emotion		
	CLIP	CLIP-H	ICAD	CLIP	CLIP-H	ICAD
SD	0.2320	0.1339	**3.64**	**0.2026**	**0.1399**	**3.84**
StruD	**0.2339**	0.1284	3.61	0.1930	0.1103	3.82
ComD	0.2299	**0.1367**	3.60	0.1901	0.1155	3.82
FairD-Gender	**0.2274**	0.1348	3.62	**0.1894**	0.1298	**3.91**
UCE-Gender	0.2280	0.1133	3.61	0.1848	1.1198	3.76
dibiasVL-Gender	0.2011	0.1223	3.47	0.1806	0.0890	3.79
FairM-Gender	0.2021	**0.1494**	**3.69**	0.1809	**0.1366**	**3.91**
FairD-Race	**0.2239**	0.1292	3.64	0.1882	0.1266	3.89
UCE-Race	0.2327	0.1045	3.58	**0.1914**	0.0968	3.77
dibiasVL-Race	0.2187	0.1022	**3.68**	0.1795	0.1083	3.77
FairM-Race	0.2197	**0.1522**	**3.68**	0.1848	**0.1324**	**3.95**
FairD-Age	0.2171	0.1084	3.56	0.1846	0.0923	3.55
UCE-Age	**0.2327**	0.1045	3.58	**0.1914**	0.0968	3.77
dibiasVL-Age	0.2187	0.0942	**3.62**	0.1758	0.0949	3.88
FairM-Age	0.2141	**0.1520**	**3.62**	0.1743	**0.1157**	**3.95**

Table 2: Evaluation results of image alignment and diversity. CLIP denotes as CLIP-Score and CLIP-H denotes as CLIP-Human. The **bold** value denotes the best performance.

Regarding diversity, for the Occupation dataset in Table 2, our method demonstrates a significant improvement in ICAD of generated results when compared to Stable Diffusion. Specifically, the metrics show an increase from 3.64 to 3.68 and 3.69 for gender and race groups, respectively. For the Emotion dataset, the ICAD of our method is still better

than other debiasing text-to-image methods. These results demonstrate our method's ability to excel in generating varied results in diverse environments.

Computation cost. On an Nvidia V100, our debiasing method trains on 150 occupations in just 50 minutes, demonstrating impressive efficiency. As shown in Table 3, our method generates 100 images for a single occupation in 434 seconds, comparable to SD. While UCE increases the minimal inference time by adjusting origin parameters, it requires significantly more training time.

Models	Wall-clock Time (seconds)
SD	424
FairD	1463
UCE	426
debiasVL	764
FairM(ours)	434

Table 3: Evaluation results in wall-clock time consumption on generation of 100 images.

Human preference. We conduct a human study about the fidelity and alignment of our method in Table 4. For fidelity, human preference scores reveal our method consistently outperforms other generative methods both for occupation and emotion description. As our method introduces a trade-off between prioritizing fairness and maintaining the alignment of facial expressions and textual descriptions, some participants expressed dissatisfaction to our method's performance in achieving consistency between facial expressions and textual descriptions. Future research may focus on enhancing the consistency of text prompts while balancing the bias.

Models	Occupation		Emotion	
	Fidelity	Alignment	Fidelity	Alignment
SD	2.7558	**3.6760**	2.7230	3.4929
StruD	2.5399	2.9953	3.3427	3.3615
ComD	2.6667	3.0375	1.9718	**3.6854**
FairM-Gender	3.0140	3.0760	3.4883	3.2431
FairM-Race	3.0798	3.3661	3.0140	3.3475
Real Image	**3.4694**	-	**3.5576**	-

Table 4: Evaluation results for Human Preference. The higher the score, the more it aligns with human preferences. Please refer to the Appendix for details.

Ablation Study

Finally, we conduct an ablation study on the necessities of our two components $\mathcal{L}_{text}$ in Eq.5 and $\mathcal{L}_{fair}$ in loss (6). Table 5 shows the results of an ablation study examining the influence of different factors in the loss terms on model performance. We implement our experiments on sensitive attribute groups Gender and dataset Occupation. First, we can see that $\mathcal{L}_{text}$ can function independently, as indicated by individual rows representing the method's performance when only one of these criteria is considered. However, it is evident that $\mathcal{L}_{fair}$ alone is not effective, indicating the necessity of $\mathcal{L}_{text}$ to establish an effective semantic space. Secondly, we can observe the combination of generating diverse sensitive attributes ($\mathcal{L}_{text}$) and maintaining fairness in representation ($\mathcal{L}_{fair}$) achieves the lowest diffusion bias, indicating the superior performance of fairness.

$\mathcal{L}_{text}$	$\mathcal{L}_{fair}$	Bias (O.)	Bias (E.)
-	-	0.4466	0.4622
-	✓	-	-
✓	-	0.4030	0.3862
✓	✓	**0.3624**	**0.2113**

Table 5: An ablation study on $\mathcal{L}_{fair}$ and $\mathcal{L}_{text}$ in the loss function. O. denotes the Occupation dataset and E. denotes the Emotion dataset.

In Figure 5, we study the effect of the fairness penalty regularization parameter λ. Firstly, we can see that when λ becomes larger, both CLIP-Score and Huma-CLIP decrease. This is due to a larger λ implying that we will more focus on fairness rather than the quality of the generated images. Moreover, these two metrics only slightly decrease when λ is less than 0.1, which further supports our previous conclusion that our method has almost the same quality of generated images. Secondly, we can observe that when λ is larger and smaller than 0.1, the BiasScore will decrease. However, when λ is greater than 0.1, the BiasScore will increase. We think in this case, the generated images experience severe distortion, resulting in a reduced amount of semantic information and consequently leading to a decline in fairness as well.

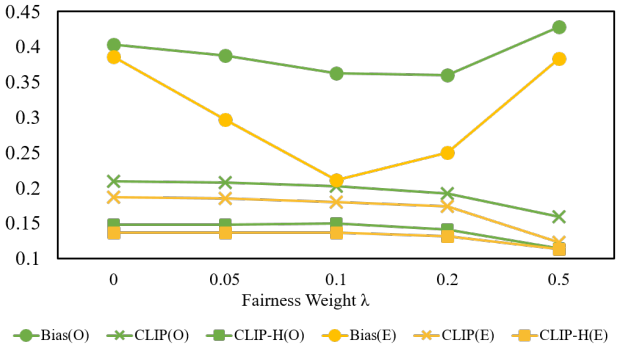

Figure 5: **The influence of different regularization parameter** λ. (O) and (E) denote Occupation dataset and Emotion dataset, respectively.

Conclusion

In this paper, we advocate that the implicit biases from input text prompts contribute to significant observed bias in current text-to-image diffusion models. We develop Fair Mapping, a model-agnostic debiasing mapping network, to effectively mitigate bias with few additional parameters for training. Meanwhile, it is flexible for Fair Mapping to adapt different customized data. Furthermore, in fairness evaluation metrics, experiments demonstrate substantial efficiency compared to text-guided diffusion models and other debiasing methods.

Acknowledgements

Di Wang and Lijie Hu are supported in part by the funding BAS/1/1689-01-01, URF/1/4663-01-01, REI/1/5232-01-01, REI/1/5332-01-01, and URF/1/5508-01-01 from KAUST, and funding from KAUST - Center of Excellence for Generative AI, under award number 5940. Hua Zhang and Jia Li are supported in part by the National Key Research and Development Program of China under Grant 2021YFB3100800; in part by the National Natural Science Foundation of China under Grant 62372448.

References

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent Anti-Muslim Bias in Large Language Models. In Fourcade, M.; Kuipers, B.; Lazar, S.; and Mulligan, D. K., eds., *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, 298–306. ACM.

Arad, D.; Orgad, H.; and Belinkov, Y. 2023. ReFACT: Updating Text-to-Image Models by Editing the Text Encoder. *arXiv preprint arXiv:2306.00738*.

Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2023. SpaText: Spatio-Textual Representation for Controllable Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 18370–18380. IEEE.

Balaji, Y.; Nah, S.; and Xun Huang, e. a. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *CoRR*, abs/2211.01324.

Bansal, H.; Yin, D.; Monajatipoor, M.; and Chang, K. 2022. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 1358–1370. Association for Computational Linguistics.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Elish, M. C.; Isaac, W.; and Zemel, R. S., eds., *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, 610–623. ACM.

Berg, H.; Hall, S. M.; and Yash Bhalgat, e. a. 2022. A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning. arXiv:2203.11933.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *CoRR*, abs/2110.01963.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. arXiv:2304.08818.

Ceylan, D.; Huang, C. P.; and Mitra, N. J. 2023. Pix2Video: Video Editing using Image Diffusion. *CoRR*, abs/2303.12688.

Chen, T.; Kornblith, S.; and Mohammad Norouzi, e. a. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *CoRR*, abs/2002.05709.

Chuang, C.-Y.; Jampani, V.; and Li, e. a., Yuanzhen. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.

Dehouche, N. 2021. Implicit Stereotypes in Pre-Trained Classifiers. *IEEE Access*, 9: 167936–167947.

Dhariwal, P.; and Nichol, A. Q. 2021. Diffusion Models Beat GANs on Image Synthesis. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *NeurIPS 2021, December 6-14, 2021, virtual*, 8780–8794.

Ding, L.; Yu, D.; Xie, J.; Guo, W.; Hu, S.; Liu, M.; Kong, L.; Dai, H.; Bao, Y.; and Jiang, B. 2021. Word Embeddings via Causal Inference: Gender Bias Reducing and Semantic Information Preserving. arXiv:2112.05194.

Dokmanic, I.; Parhizkar, R.; Ranieri, J.; and Vetterli, M. 2015. Euclidean Distance Matrices: Essential theory, algorithms, and applications. *IEEE Signal Process. Mag.*, 32(6): 12–30.

El-Nouby, A.; Sharma, S.; Schulz, H.; and R. Devon Hjelm, e. a. 2018. Keep Drawing It: Iterative language-based image generation and editing. *CoRR*, abs/1811.09845.

Feng, W.; He, X.; and Tsu-Jui, e. a. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Feng, Z.; Zhang, Z.; Yu, X.; Fang, Y.; Li, L.; Chen, X.; Lu, Y.; Liu, J.; Yin, W.; Feng, S.; Sun, Y.; Chen, L.; Tian, H.; Wu, H.; and Wang, H. 2023. ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts. arXiv:2210.15257.

Fraser, K. C.; Kiritchenko, S.; and Nejadgholi, I. 2023. A Friendly Face: Do Text-to-Image Systems Rely on Stereotypes when the Input is Under-Specified? *arXiv preprint arXiv:2302.07159*.

Friedrich, F.; Schramowski, P.; and Manuel Brack, e. a. 2023. Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. *CoRR*, abs/2302.10893.

Gandikota, R.; Orgad, H.; and Belinkov, e. a., Yonatan. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF WACV*, 5111–5120.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, volume EMNLP 2020 of *Findings of ACL*, 3356–3369. Association for Computational Linguistics.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413.

He, Y.; Salakhutdinov, R.; and Kolter, J. Z. 2023. Localized Text-to-Image Generation for Free via Cross Attention Control. *CoRR*, abs/2306.14636.

Hessel, J.; Holtzman, A.; and Maxwell Forbes, e. a. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 7514–7528. Association for Computational Linguistics.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; and Raia Hadsell, e. a., eds., *NeurIPS 2020, 6-12, 2020, virtual*.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *CoRR*, abs/2207.12598.

Jiang, Y.; Lyu, Y.; and Ma, e. a. 2023. RS-Corrector: Correcting the Racial Stereotypes in Latent Diffusion Models. *arXiv preprint arXiv:2312.04810*.

Kärkkäinen, K.; and Joo, J. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, 1547–1557. IEEE.

Karras, T.; Laine, S.; and Aila, T. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. *CoRR*, abs/1812.04948.

Kim, Y.; Lee, J.; Kim, J.; Ha, J.; and Zhu, J. 2023. Dense Text-to-Image Generation with Attention Modulation. *CoRR*, abs/2308.12964.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Bengio, Y.; and LeCun, Y., eds., *ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Le, N.; and Odobez, J. 2018. Robust and Discriminative Speaker Embedding via Intra-Class Distance Variance Regularization. In Yegnanarayana, B., ed., *Interspeech 2018, Hyderabad, India, 2-6 September 2018*, 2257–2261. ISCA.

Makady, A.; de Boer, A.; Hillege, H.; Klungel, O.; and Goettsch, W. 2017. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. *Value in Health*, 20(7): 858–865.

Meng, K.; Sharma, A. S.; and Andonian, e. a., Alex. 2022. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Mikolov, T.; Chen, K.; and Greg Corrado, e. a. 2013. Efficient Estimation of Word Representations in Vector Space. In Bengio, Y.; and LeCun, Y., eds., *ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Ning, M.; Li, M.; and Jianlin Su, e. a. 2023. Elucidating the Exposure Bias in Diffusion Models. *CoRR*, abs/2308.15321.

Orgad, H.; Kawar, B.; and Belinkov, Y. 2023. Editing implicit assumptions in text-to-image diffusion models. *arXiv preprint arXiv:2303.08084*.

Otani, M.; Togashi, R.; and Yu Sawai, e. a. 2023. Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. In *CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 14277–14286. IEEE.

Ouyang, Y.; Xie, L.; and Cheng, G. 2022. Improving Adversarial Robustness by Contrastive Guided Diffusion Process. *CoRR*, abs/2210.09643.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Radford, A.; Kim, J. W.; and Chris Hallacy, e. a. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; and Alex Nichol, e. a. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR*, abs/2204.06125.

Rombach, R.; Blattmann, A.; and Dominik Lorenz, e. a. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR 2022*, 10674–10685. IEEE.

Saharia, C.; Chan, W.; and Saurabh Saxena, e. a. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *2023 (CVPR)*, 22522–22531.

Schuhmann, C.; Beaumont, R.; and Richard Vencu, e. a. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Schuhmann, C.; Vencu, R.; and Romain Beaumont, e. a. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *CoRR*, abs/2111.02114.

Shen, X.; Du, C.; and Pang, e. a., Tianyu. 2023. Finetuning Text-to-Image Diffusion Models for Fairness. *arXiv preprint arXiv:2311.07604*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *9th ICLR, 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Struppek, L.; Hintersdorf, D.; and Kersting, K. 2022. The Biased Artist: Exploiting Cultural Biases via Homoglyphs in Text-Guided Image Generation Models. *CoRR*, abs/2209.08891.

Tang, Z.; Yang, Z.; and Chenguang, e. a. 2023. Any-to-Any Generation via Composable Diffusion. *CoRR*, abs/2305.11846.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tian, L.; Lai, C.; and Moore, J. D. 2018. Polarity and Intensity: the Two Aspects of Sentiment Analysis. *CoRR*, abs/1807.01466.

Wang, F.-E.; Wang, C.-Y.; Sun, M.; and Lai, S.-H. 2022. MixFairFace: Towards Ultimate Fairness via MixFair Adapter in Face Recognition. arXiv:2211.15181.

Wang, J.; Zhang, Y.; and Sang, J. 2022. FairCLIP: Social Bias Elimination based on Attribute Prototype Learning and Representation Neutralization. *CoRR*, abs/2210.14562.

Wu, Y.; Yu, N.; Li, Z.; Backes, M.; and Zhang, Y. 2022. Membership Inference Attacks Against Text-to-image Generation Models. *CoRR*, abs/2210.00968.

Zhang, Y.; Wang, J.; and Sang, J. 2022. Counterfactually Measuring and Eliminating Social Bias in Vision-Language Pre-training Models. In Magalhães, J.; Bimbo, A. D.; Satoh, S.; Sebe, N.; Alameda-Pineda, X.; Jin, Q.; Oria, V.; and Toni, L., eds., *MM 2022, Portugal, October 10 - 14*, 4996–5004. ACM.