Bias Begets Bias: the Impact of Biased Embeddings on Diffusion Models

Sahil Kuchlous [*1] **Marvin Li** [*1] **Jeffrey G. Wang** [*1]

Abstract

With the growing adoption of Text-to-Image (TTI) systems, the social biases of these models have come under increased scrutiny. Previous approaches only identify such biases and fail to diagnose their sources. In this paper, we conduct a systematic investigation of one such source: embedding spaces. First, we demonstrate theoretically and empirically that an unbiased text embedding for the input prompt is a *necessary* condition for representationally balanced diffusion models. Next, we investigate the impact of biased embeddings on the alignment of images and prompts, a common technique for evaluating diffusion models. We find that biased multimodal embeddings like CLIP result in lower alignment scores for representationally balanced TTI models, thus rewarding unfair behaviour. Finally, we develop a theoretical framework through which biases in alignment evaluation can be studied and propose bias mitigation methods. By specifically adapting the perspective of embedding spaces, we establish new fairness conditions for diffusion model generation and evaluation.

1. Introduction

As Text-To-Image (TTI) models become increasingly adept at generating complex and realistic images, they are being integrated into a wide range of commercial and creative services (Ramesh et al., 2021; Rombach et al., 2022; Saharia et al., 2022). The proliferation of these models is evident in various industries; for example, advertising agencies use them to quickly generate visual content for campaigns, while film and game developers employ them to design detailed backgrounds and characters. The ubiquity of these models, however, has also raised significant ethical and fairness concerns, from their potential for misuse to the biases they encode. Addressing these issues is critical to developing TTI models that are not only technologically advanced but also socially responsible and inclusive.

Here, we specifically study diffusion models, a class of models that form the centerpiece of most frontier TTI systems. We first consider *direct representational harm*; that is, breaches of fairness that arise from generations that are imbalanced across different protected classes. In the literature, it is generally "well-established" that most diffusion models have imbalanced representation in their generations. For instance, Perera & Patel (2023) show that popular TTI models consistently underrepresent minorities in most professions, a finding corroborated by several other papers (Wang et al., 2023; Luccioni et al., 2023; Wan et al., 2024).

Violations of fairness also arise indirectly; most prominently, a model that generates lower quality images of one group versus another creates *indirect representational harm*. To audit this type of harm, one important evaluation criteria on which TTI models are often benchmarked is the alignment/faithfulness of images to prompts—how well the contents of the generated image match the prompt. If generated images are well-aligned with the prompt across classes, then the model treats these groups equally. While some papers attempt to use human ratings for benchmarking (Lee et al., 2023), this approach is expensive and hard to scale; thus, more recent work has explored automating alignment evaluation. One method growing in popularity is CLIPScore, based on the text-image embedding CLIP (Radford et al., 2021), which measures the cosine similarity between a prompt and an image in the multimodal latent space (Hessel et al., 2022).

Embeddings play a critical role in generating from and evaluating diffusion models. Beyond pretraining data, the text embedding of the prompt is the only input to a diffusion model; similarly, many popular approaches for prompt-image alignment leverage a latent text-image space. Despite their critical role, however, few papers focus on connecting properties of embedding spaces to downstream fairness criteria. In this paper, we demonstrate that two intuitive conclusions hold, both theoretically and empirically. In Section 4, we show that biased embeddings cause biased generations. In Sec-

[*]Equal contribution [1]Harvard School of Engineering and Applied Sciences, Boston, USA. Correspondence to: Marvin Li <marvinli@college.harvard.edu>, Sahil Kuchlous <sahilkuchlous@gmail.com>, Jeffrey Wang <jgwang@college.harvard.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

tion 5, we show biased image-text embedding spaces lead to biased evaluation of diffusion (and any text-to-image) models. As modern machine learning systems increasingly integrate many components, like embeddings, our work highlights how ensuring the fairness of each part is critical for ensuring the fairness of the whole.

2. Related Work

Theory of Algorithmic Fairness. Existing work in algorithm fairness has concentrated on the setting of supervised learning, where individuals are mapped to outcomes. Traditionally, group fairness notions (like statistical parity) enforce some measure of average equal treatment between members of protected classes, whereas individual fairness notions require similar individuals to be treated similarly, under some task-specific metric (Dwork et al., 2011). More recently, *multi-group* fairness has emerged as a middle ground, which enforces fairness constraints on (up to exponentially many) subgroups within the dataset. One such notion is multicalibration (Hebert-Johnson et al., 2018), on which we base our work in Section 5, although others have also been proposed (Kearns et al., 2018).

Theory of Diffusion Models. In brief, diffusion models "learn" how to sample from some distribution of images. They do so by learning how much noise gets added to images over time so images can be sampled from that distribution by *de*noising a sample of pure Gaussian noise; the denoising function at time t is the same as the statistical *score* $\nabla \ln q_t(x)$ of the noised distribution $q_t(x)$ (Song et al., 2020). Mathematically, the convergence to the true distribution of the sampling process with a sufficiently good approximation of the score can be proved with Girsanov's thoerem (Theorem A.2); empirically, the process is implemented via a discrete-time approximation, where we iteratively denoise images over small time steps. We defer a full theoretical coverage of diffusion models to Appendix A, and note that we later leverage Theorem A.2 for our proof that biased embeddings cause biased generations.

Bias in Embedding Spaces. Bolukbasi et al. (2016) was the first to demonstrate bias present in word embeddings on the basis of gender, and proposed methods to debias such embeddings, followed by similar work on bias in race (Manzini et al., 2019; Dev & Phillips, 2019). Papakyriakopoulos et al. (2020) train a sentiment classifier on a biased word embedding and illustrate that the downstream classifier's outputs reflect the direction of biases in the input embedding. These biases have also been observed in multimodal embeddings like CLIP (Wang et al., 2021; Berg et al., 2022), and the impact of these biases on evaluating image captioning methods has been studied (Qiu et al., 2023). However, to our knowledge, our work is the first to analyze how bias in learned embedding spaces like CLIP affects the evaluation of TTI systems, as well as the downstream impact of a biased embedding as a component of a *generative* model.

Bias in Diffusion Models. Several recent papers on fairness in diffusion models propose methods to sample with more equal representation across protected classes (Li et al., 2024; Friedrich et al., 2023; Choi et al., 2024; Shen et al., 2024; Chuang et al., 2023). Of note, (Chuang et al., 2023) and (Li et al., 2024) intervene on the prompt embedding, using a debiasing projection and a learned fair representation of prompts, respectively. Both works illustrate that debiasing embeddings leads to more representationally fair generations. In Section 4, we show this is a necessary, and not simply a sufficient, condition.

Alignment. There are several papers that attempt to benchmark the prompt-image alignment of image generation models. Of these, Lee et al. (2023) rely most heavily on CLIPScore for alignment evaluation. Bakr et al. (2023) use CLIPScore for more specific evaluation criteria like generation of emotion. Chen et al. (2024) define a composite metric called 'text condition evaluation' that encompasses alignment and fairness. To calculate the alignment portion of their score, the authors utilize a visual question answering (VQA) model called BLIP (Li et al., 2022). Bakr et al. (2023) and Chen et al. (2024) also raise concerns with utilizing CLIPScore to measure alignment, but do not cite fairness as a concern. There have been several other attempts at evaluating text-image alignment (Hu et al., 2023; Xu et al., 2023; Yarom et al., 2023), all of which potentially demonstrate some form of implicit bias. In this paper, we introduce a framework for studying these biases irrespective of the underlying alignment score function used.

3. Preliminaries

Here, we introduce some notation that we will use throughout the manuscript. Let TV denote the total variation distance. Let $\mathcal{P}$ denote the set of all prompts (e.g. 'an image of a doctor interacting with a patient') and $\mathcal{I}$ to denote the set of all images. A text-to-image model $M : \mathcal{P} \to \Delta(\mathcal{I})$ takes a prompt as the input and returns a distribution over images that can be sampled from. A multimodal embedding consists of a pair of functions $(e_{\mathcal{I}} : \mathcal{I} \to \mathcal{S}^{n-1}, e_{\mathcal{P}} : \mathcal{P} \to \mathcal{S}^{n-1})$, where $\mathcal{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$. We use the set $\mathcal{A} = \{a_1, ..., a_k\}$ to refer to possible attributes such as race or gender and $b \in \mathcal{P}$ to denote the base prompt (e.g. "firefighter," or more generally any descriptor independent of $\mathcal{A}$). For some $a \in \mathcal{A}$, the operation $a + b \in \mathcal{P}$ represents the prompt obtained by combining an attribute and a base prompt. For example, for $a =$ 'male' and $b =$ 'doctor', we then have $a + b =$ 'male doctor'. We also use $e_{\mathcal{P}}$ to denote the map from prompts to prompt embeddings within a diffusion model. Let $p_y : \mathcal{I} \to \mathbb{R}^{\geq 0}$ be the distribution over images generated by the model conditioned on text

prompt y.

4. Biased Embedding, Biased Generations

In Theorem 4.2, we prove that sufficiently strong bias in the prompt's embeddings implies bias in the image generations. We begin by providing some intuition to interpret the result. Suppose we have a base prompt $b =$ 'nurse' and attributes $a_1 =$ 'man' and $a_2 =$ 'woman'. If the embedding of b is sufficiently close to the embedding for $a_2 + b$ and images with the attributes a_1 and a_2 are sufficiently distinguished, Theorem 4.2 states that the diffusion model conditioned on b will mostly produce images similar to when it is conditioned on $a_2 + b$. Assuming that the images generated by conditioning on $a_2 + b$ are faithful to the prompt, most of the images generated from 'nurses' would be of 'woman nurses', which is exactly representational bias in sampling.

We begin with some definitions. Let $s_t(x, y)$ denote the learned score function of a diffusion model at input x and time t conditioned on a prompt embedding y. Let T be the length of the denoising period. We consider the bias in a model's output with a base prompt b and set of attributes $\{a_1, a_2, \ldots, a_k\}$, which should be thought of as distinct categories, and consider a small positive $\epsilon \in (0, 1/2)$.

Assumption 4.1 (Lipschitz in prompt embeddings)**.** The score $s_t(x, y)$ of the diffusion model is L-Lipschitz in y.

Assumption 4.1 states that the score function is Lipschitz with respect to the prompt embeddings and is key to Theorem 4.2. [1] Assumption 4.1 can also be interpreted as an *individual fairness constraint*, where the requirement that similar prompt embeddings produces similar images parallels the notion that similar individuals are mapped to similar outcomes (Dwork et al., 2011). With this machinery in hand, we state our main theorem below, which formalizes the intuition that embeddings where a base prompt is close to the prompt + attribute gives a bound on the distribution of images generated. The result below leverages a key result from Chen et al. (2023), which we use as a black-box here; see Theorem A.2 in the appendix for more details.

Theorem 4.2 (Bias in embeddings implies bias in image generations)**.** *Assume there exists $\epsilon \in (0, 1/2)$ such that for all distinct $a_i, a_j \in \mathcal{A}$, $\mathrm{TV}(p_{a_i+b}, p_{a_j+b}) \geq 1 - \epsilon$. Without loss of generality assume that a_1 minimizes the distance with b, i.e. $\mathrm{argmin}_{j=1,\ldots,k} \|e_{\mathcal{P}}(b) - e_{\mathcal{P}}(a_j + b)\| = 1$, and let $\|e_{\mathcal{P}}(b) - e_{\mathcal{P}}(a_1 + b)\| \leq \frac{\epsilon}{\sqrt{TL}}$. Under Assumption 4.1, we have $\mathrm{TV}(p_b, p_{a_1+b}) \leq \epsilon$ and $\mathrm{TV}(p_b, p_{a_j+b}) \geq 1 - 2\epsilon$ for $j \neq 1$.*

Proof. We begin by noting that the second claim in the theorem, $\mathrm{TV}(p_b, p_{a_j+b}) \geq 1 - 2\epsilon$ for $j \neq 1$, follows immediately from $\mathrm{TV}(p_{a_i+b}, p_{a_j+b}) \geq 1 - \epsilon$ for all distinct $i, j = 1, 2, \ldots, k$, the triangle inequality, and the first claim of the theorem. Hence, we only need to show that $\mathrm{TV}(p_b, p_{a_1+b}) \leq \epsilon$.

By Assumption 4.1, we have an upper bound $\|s_t(x, e_{\mathcal{P}}(b)) - s_t(x, e_{\mathcal{P}}(a_1 + b))\| \leq L\|e_{\mathcal{P}}(a_1 + b) - e_{\mathcal{P}}(b)\|$. Theorem A.2 translates this into a bound on $\mathrm{KL}(p_b, p_{a_1+b})$, and we finish by applying Pinsker's inequality, which gives us that

$$\mathrm{TV}(p_b, p_{a_1+b}) \lesssim \sqrt{TL^2 \mathbb{E}_x \|e_{\mathcal{P}}(a_1 + b) - e_{\mathcal{P}}(b)\|^2} \leq \epsilon.$$

□

In the theorem above, we see that the distribution of images when the diffusion model is conditioned on b is similar to the distribution when one conditions on $a_1 + b$ and far from when one conditions on $a_j + b$ for $j \neq 1$. Note the requirement that $\|e_{\mathcal{P}}(b) - e_{\mathcal{P}}(a_1 + b)\| \leq \frac{\epsilon}{\sqrt{TL}}$ is difficult to realize in practice as the embedding distance has to be extremely small to meaningfully control the total variation.

Empirics. Previous work has illustrated that the output of diffusion models are not representationally balanced. For instance, Friedrich et al. (2023) illustrate gender bias in the outputs of occupations queried of Stable Diffusion which reflect both the direction of biases in CLIP embeddings and representational imbalances in the underlying training dataset. We find a similar relationship when comparing generations across occupations from Stable Diffusion 2.1 (SD2.1) against biases in the underlying CLIP embedding (Appendix B).

Since SD2.1 is trained on an underlying dataset (LAION-5B) that is itself representationally imbalanced, however, it is unclear if the bias in image outputs is caused by a biased prompt embedding or imbalanced training dataset. To probe this, we train a conditional diffusion model from scratch with balanced training data across three classes (nurse, philosopher, and person) but with biased prompt embeddings (where nurse is closer to woman, philosopher is closer to man, and person is roughly equidistant). We find that the resulting model *is* biased in its generations in the same direction as the embedding, with women as the majority of nurse generations, men as the majority of philosopher generations, and a roughly even split in generations of people. We defer the full experimental details to Appendix B.2.

5. Bias in Alignment Auditing

In this section we consider the problem of auditing image generation models for prompt-image alignment and analyze the impact that biased multimodal embeddings may have

[1] Lipschitzness of the learned score function in the first argument is standard in the theory of diffusion literature, e.g., Chen et al. (2023)

on the fairness of alignment scores. We note that from here on out, we refer to "score" as the alignment score, not the learned function of a diffusion model. While the results in Section 4 studies the representational harm caused by biased diffusion models, this section analyzes harms caused by biased alignment predictors and therefore borrows from richer notions of fairness proposed for classification systems. We define a notion of fairness for alignment functions and demonstrate *necessary* conditions for multimodal embeddings to satisfy this definition. Additionally, we evaluate the bias of existing auditing functions and suggest simple techniques for mitigating such biases.

5.1. Definitions of Fairness

We begin by defining what it means for an alignment auditing function to be fair. These definitions are inspired by existing work on fairness for predictors introduced by (Hebert-Johnson et al., 2018), and are designed to capture the idea that the alignment of an image with a prompt should be independent of protected attributes like race and gender when they are not explicitly specified in the prompt. We use $s^* : \mathcal{P} \times \mathcal{I} \to [0, 1]$ to denote the true alignment score and $s : \mathcal{P} \times \mathcal{I} \to [0, 1]$ to denote an auditing function.

Definition 5.1 (Multiaccuracy)**.** Let $\mathcal{C} \subseteq 2^{\mathcal{I}}$ be a collection of subsets of $\mathcal{I}$ and $\alpha \in [0, 1]$. An auditing function s is $(\mathcal{C}, \alpha)$-multiaccurate for prompt $b \in \mathcal{P}$ if, for all $I \in \mathcal{C}$,

$$\left| \mathop{\mathbb{E}}_{i \sim I}[s^*(b, i) - s(b, i)] \right| \leq \alpha.$$

In other words, multiaccuracy for a prompt b ensures that the function s is α-accurate on every subset of images $I \in \mathcal{C}$. To see how this notion is useful for fairness, consider a set of protected attributes $a_1, ..., a_k$ for a base prompt b, and let I_ℓ be a subset of images corresponding to prompt b with attribute a_ℓ. If $\mathcal{C} = \{I_1, ..., I_k\}$, $(\mathcal{C}, \alpha)$-multiaccuracy guarantees that the auditing function s is α-accurate on average on each of the protected attributes. Thus, for example, if the prompt we care about is 'doctor' and the attributes we care about encompass gender, setting I_1 and I_2 to images of male and female doctors respectively gives us the guarantee that s is α-accurate on both genders.[2] We analyze the strengths and weaknesses of multiaccuracy and define a stronger fairness guarantee based on multicalibration in Appendix C.1.

To see why these notions of fairness are useful for auditing text-to-image models, we prove the following theorem. Intuitively, this theorem states that if we have subsets of images with particular attributes such that the true alignment is balanced across attributes, and if the auditor is multiaccurate on these subsets of images, then the alignment score remains stable irrespective of how a text-to-image model chooses to sample from these attributes.

Theorem 5.2. *Consider a base prompt $b \in \mathcal{P}$ and attributes $\mathcal{A} = \{a_1, ..., a_k\}$, and let I_ℓ be a subset of images corresponding to prompt b with attribute a_ℓ. Assume that $\mathbb{E}_{i \sim I_\ell}[s^*(b, i)] = \bar{s}$ for all $\ell \in [k]$. Let $\mathcal{C} = \{I_1, ..., I_k\}$. Consider a model M that, given the prompt b, returns an image $i \sim I_\ell$ with probability p_ℓ, where $\sum_{\ell \in [k]} p_\ell = 1$. If the auditing function s is $(\mathcal{C}, \alpha)$-multiaccurate, then $\left|\mathbb{E}_{i \sim M(b)}[s(b, i)] - \bar{s}\right| \leq \alpha$ irrespective of the probabilities $p_1, ..., p_k$.*

Proof. We see that

$$\begin{aligned}
\mathop{\mathbb{E}}_{i \sim M(b)}[s(b, i)] &= \sum_{\ell \in [k]} \mathop{\mathbb{E}}_{i \sim I_\ell}[s(b, i)] \cdot p_\ell \\
&\leq \sum_{\ell \in [k]} \left(\mathop{\mathbb{E}}_{i \sim I_\ell}[s^*(b, i)] + \alpha \right) \cdot p_\ell \\
&= \sum_{\ell \in [k]} (\bar{s} + \alpha) \cdot p_\ell \\
&= \bar{s} + \alpha.
\end{aligned}$$

By the same logic, we also see that $\mathbb{E}_{i \sim M(b)}[s(b, i)] \geq \bar{s} - \alpha$. Thus, $-\alpha \leq \mathbb{E}_{i \sim M(b)}[s(b, i)] - s \leq \alpha$, so $\left|\mathbb{E}_{i \sim M(b)}[s(b, i)] - \bar{s}\right| \leq \alpha$. □

As a corollary of this theorem, we see that irrespective of the probabilities $p_1, ..., p_n$, the difference in the alignment score of two models is at most 2α. Thus, going back to our example of male and female doctors, we see that as long as we expect the average alignment of an image of a male doctor and a female doctor with the prompt 'doctor' to be the same, if our auditing function is multiaccurate on the appropriate subsets of images, models will get similar alignment scores irrespective of what proportion of male or female doctors they generate.

Finally, it is worth noting that checking whether an auditing function is multiaccurate may be challenging, since this requires access to the true alignment scores s^*. Note that if we had arbitrary oracle access to s^*, there would be no need to evaluate auditing functions s at all, since we could just use s^* to audit the alignment of text-to-image models. However, the advantage of our notion of fairness is that it only requires 'true' scores for a diverse but fixed set of images on which auditing functions can then be evaluated. Such datasets could be obtained by manual labeling, and there have been several efforts to create large image-caption datasets with alignment ratings (Levinboim et al., 2021; Lee et al., 2021; Vedantam et al., 2015; Hodosh et al., 2013).

[2] Note that gender is not a binary, and we would hope that generative models are able to capture gender on a spectrum. We use male and female in the example for ease of notation.

5.2. Properties of Fair Embeddings

Let us begin by defining an auditing function in terms of a multi-modal embedding space. For vectors $\mathbf{x}$ and $\mathbf{y}$, we define their cosine similarity as $\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$. Given a multimodal embedding $(e_{\mathcal{I}}, e_{\mathcal{P}})$, we define the auditing function $s(b, i) = \frac{\cos\left(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i)\right)+1}{2}$. This ensures that $s(b, i) \in [0, 1]$. Note that existing techniques like CLIPScore (Hessel et al., 2022) achieve a similar effect by clipping the cosine similarity at a minimum of 0, but this is equivalent to our definition up to a factor of 2 as long as similarity is positive.

Given the definitions of fairness for alignment auditors in Section 5.1, it is natural to ask what properties we would expect from a multimodal embedding space in order to obtain a fair auditor. We start with the following observation, which provides a necessary condition for multiaccuracy based on the average score between a prompt and different subsets of images.

Theorem 5.3. *If s is $(\mathcal{C}, \alpha)$-multiaccurate for a prompt b, for all $I_\ell, I_{\ell'} \in \mathcal{C}$, if $\mathbb{E}_{i \in I_\ell}[s^*(b, i)] = \mathbb{E}_{i \in I_{\ell'}}[s^*(b, i)]$, then $|\mathbb{E}_{i \in I_\ell}[\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i))] - \mathbb{E}_{i \in I_{\ell'}}[\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i))]| \leq 4\alpha$.*

Proof. Since s is $(\mathcal{C}, \alpha)$-multiaccurate for prompt b, for all $I \in \mathcal{C}$ we know that

$$\left|\mathop{\mathbb{E}}_{i \sim I}[s^*(b, i) - s(b, i)]\right| = \left|\mathop{\mathbb{E}}_{i \sim I}[s^*(b, i)] - \mathop{\mathbb{E}}_{i \sim I}[s(b, i)]\right| \leq \alpha.$$

Moreover, for some $I_\ell, I_{\ell'} \in \mathcal{C}$, if $\mathbb{E}_{i \in I_\ell}[s^*(b, i)] = \mathbb{E}_{i \in I_{\ell'}}[s^*(b, i)]$, by the triangle inequality we see that

$$\left|\mathop{\mathbb{E}}_{i \sim I_\ell}[s(b, i)] - \mathop{\mathbb{E}}_{i \sim I_{\ell'}}[s(b, i)]\right| \leq 2\alpha.$$

Substituting the definition of $s(b, i)$, we see that

$$\left|\mathop{\mathbb{E}}_{i \sim I_\ell}[\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i))] - \mathop{\mathbb{E}}_{i \sim I_{\ell'}}[\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i))]\right| \leq 4\alpha.$$

□

This implies that given a prompt b and subsets of images with similar average true alignment scores, if the embeddings of those images do not have a similar distance (on average) to the embedding of b, then our auditor is not multiaccurate on those images.

Next, we use this observation to detect bias based only on the prompt embeddings $e_{\mathcal{P}}$. As discussed in Section 4, prompt embeddings are often biased in themselves, which causes bias in diffusion models when these embeddings are used as inputs. By showing that fair prompt embeddings are necessary for multiaccuracy, the theorem below demonstrates that bias in prompt embeddings also translates to bias in multimodal embeddings, and therefore results in biased alignment auditing.

Theorem 5.4. *Consider a prompt b and attributes $\mathcal{A} = \{a_1, ..., a_k\}$. Let I_ℓ be a set of images such that for every $i \in I_\ell$, $e_{\mathcal{I}}(i)$ is in a ball of radius ε around $e_{\mathcal{P}}(a_\ell + b)$. For $\mathcal{C} = \{I_1, ..., I_k\}$, if s is $(\mathcal{C}, \alpha)$-multiaccurate for prompt b, for any $\ell, \ell' \in [k]$ such that $\mathbb{E}_{i \in I_\ell}[s^*(b, i)] = \mathbb{E}_{i \in I_{\ell'}}[s^*(b, i)]$, it holds that $|\cos(e_{\mathcal{P}}(b), e_{\mathcal{P}}(a_\ell + b)) - \cos(e_{\mathcal{P}}(b), e_{\mathcal{P}}(a_{\ell'} + b))| \leq 4\alpha + 2\varepsilon$.*

Proof. Since s is $(\mathcal{C}, \alpha)$-multiaccurate for prompt b, if $\mathbb{E}_{i \in I_\ell}[s^*(b, i)] = \mathbb{E}_{i \in I_{\ell'}}[s^*(b, i)]$, we know by Theorem 5.3 that

$$\left|\mathop{\mathbb{E}}_{i \sim I_\ell}[\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i))] - \mathop{\mathbb{E}}_{i \sim I_{\ell'}}[\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i))]\right| \leq 4\alpha.$$

Next, consider the expression $\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i))$. For all $i \in I_\ell$, since $e_{\mathcal{I}}(i)$ is in a ball of radius ε around $e_{\mathcal{P}}(a_\ell + b)$, we know that there exists a vector δ such that $e_{\mathcal{I}}(i) = e_{\mathcal{P}}(a_\ell + b) + \delta$ and $\|\delta\| \leq \varepsilon$. Thus, we see that

$$\begin{aligned}
\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i)) &= \cos(e_{\mathcal{P}}(b), e_{\mathcal{P}}(a_\ell + b) + \delta) \\
&= \frac{e_{\mathcal{P}}(b) \cdot (e_{\mathcal{P}}(a_\ell + b) + \delta)}{\|e_{\mathcal{P}}(b)\|\|e_{\mathcal{P}}(a_\ell + b) + \delta\|} \\
&= e_{\mathcal{P}}(b) \cdot e_{\mathcal{P}}(a_\ell + b) + e_{\mathcal{P}}(b) \cdot \delta \\
&= \cos(e_{\mathcal{P}}(b), e_{\mathcal{P}}(a_\ell + b)) + e_{\mathcal{P}}(b) \cdot \delta.
\end{aligned}$$

Since $\|\delta\| \leq \varepsilon$, we know that $|e_{\mathcal{P}}(b) \cdot \delta| \leq \varepsilon$. Thus,

$$|\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i)) - \cos(e_{\mathcal{P}}(b), e_{\mathcal{P}}(a_\ell + b))| \leq \varepsilon,$$

for all $i \in I_\ell$, which implies that

$$\left|\mathop{\mathbb{E}}_{i \sim I_\ell}[\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i))] - \cos(e_{\mathcal{P}}(b), e_{\mathcal{P}}(a_\ell + b))\right| \leq \varepsilon.$$

By symmetry, note that the same expression holds for $i \sim I_{\ell'}$ and $a_{\ell'}$. Thus, by the triangle inequality,

$$|\cos(e_{\mathcal{P}}(b), e_{\mathcal{P}}(a_\ell + b)) - \cos(e_{\mathcal{P}}(b), e_{\mathcal{P}}(a_{\ell'} + b))| \leq 4\alpha + 2\varepsilon.$$

□

The contrapositive of this theorem implies that, for a prompt b, if there are two attributes a_1 and a_2 such that the embeddings of $a_1 + b$ and $a_2 + b$ are not equidistant from the embedding of b, then the resultant auditing function s can not be fair on images that closely match attributes a_1 and a_2. Thus, if the underlying embedding space is not fair on prompts, we should not expect it to be a fair auditor! This allows us to infer bias in a multimodal embedding space by only looking at its prompt embedding function $e_{\mathcal{P}}$.

5.3. Auditing Alignment with Biased Embeddings

While there have been some efforts to debias CLIP and CLIPScore (Dehdashtian et al., 2024), it is likely that many of the approaches we use to audit prompt-image alignment will continue to exhibit some form of representational bias. Thus, it is important to consider methods to mitigate this bias in the score calculation phase. Currently, the standard method for calculating the alignment score of a model M for a prompt b using a multimodal embedding is to take the average of the auditing function described in Section 5.2 over several images samples from $M(b)$ (Lee et al., 2023). Henceforth, we will refer to this method as *score-then-average*. As we have discussed, one of the ways in which an embedding space may be biased is if for a prompt b, vectors corresponding to images with attribute a_1 are more similar to the vector for b on average than those with attribute a_2. In this case, a model that generates only images with attribute a_1 will consistently get a higher alignment score than a model that generates a balanced distribution of images. In this section, we suggest an alternative method of evaluating alignment based on potentially biased multimodal embeddings that helps alleviate this form of bias. We call this method *subclass-score*.

In this approach, given a prompt b, we begin with a list of attributes $a_1, ..., a_k$ on which we would like to be fair. To calculate the alignment score of a model M of this prompt, we start by sampling a set I of images from $M(b)$. We then calculate the individual scores of images $i \in I$ as

$$s(b, i) = \max_{\ell \in [k]} (s(a_k + b, i)).$$

For example, if $a_1 =$ 'male' and $b =$ 'doctor', $a_1 + b =$ 'male doctor'. Finally, to calculate the aggregate score we take the average of $s(b, i)$ over images $i \in I$.

This aims to tackle the issue where embeddings of images with one attribute may be closer on average to the embedding of the prompt than images with other attributes. By measuring the similarity with attribute-specific vectors instead, the relative distance from the embedding of the original prompt is no longer an issue. We note that there remain drawbacks with this approach. First, it may be difficult to identify what attributes are appropriate for which prompts. For instance, Google's Gemini model recently attempted to force diversity into prompts where it was not historically accurate or appropriate (Gautam et al., 2024). Also, this does not guarantee fairness conditions like multiaccuracy. For example, it is possible that images with some attributes a_ℓ are closer on average to the prompt $a_\ell + b$ than other attributes $a_{\ell'}$ are to the prompt $a_{\ell'} + b$.

We introduce a second method for generating unbiased alignment scores, *average-then-score*, in Appendix C.2. This method involves taking the mean of the image embeddings before computing the cosine similarity to the prompt embedding. We evaluate and compare these approaches in Section 5.4.

5.4. Empirical Case-Study: CLIP

In this section, we investigate biases present in an existing method of alignment auditing called CLIPScore (Hessel et al., 2022) based on the multimodal embedding model CLIP (Radford et al., 2021). We then evaluate the techniques proposed in Section 5.3. We find that the text embeddings of CLIP demonstrate gender bias in the representation of occupations, matching the findings of (Bolukbasi et al., 2016). We also find that a similar bias is present in the relationship between text embeddings of occupations and gendered image embeddings. In particular, we see that for the same set of images, male medical professionals got a 0.02 higher score than female medical professionals on average for the prompt 'doctor,' and a 0.061 lower score for the prompt 'nurse.' This clearly demonstrates an underlying bias in the auditing CLIPScore function; models that generate only male doctors and only female nurses get over 5% higher alignment scores than models with balanced distributions. Finally, we evaluate our bias mitigation methods in Section 5.3 and find that subclass-score performs significantly better than average-then-score, staying roughly consistent across gender ratio. Thus, subclass-score is the most promising step for alleviating gender bias in alignment scores for text-to-image models. Further details can be found in Appendix C.3.3.

6. Conclusion and Future Work

In this paper we study the relationship between bias in embeddings and image generations for text-to-image models. We prove theoretically and empirically that biased prompt embeddings lead to representationally unfair outputs. This establishes a *necessary* condition (an unbiased embedding) for an unbiased model. We then investigate the impact of biased prompt embeddings on measuring the alignment between image generations and prompts, as well as multi-accuracy and multi-calibration style definitions for fair alignment auditing algorithms.

There are several interesting directions to take this work. First, the text embedding's impact on fairness in diffusion models can be further investigated empirically with recently-released open state-of-the-art models that have public training sets (Gokaslan et al., 2023). Next, we are interested in potential augmentations for training diffusion models that would compensate for bias in embedding spaces. Finally, while we give several conditions that imply a metric-based auditing function is not fair, it would be interesting to to explore *sufficient* conditions for fairness, like in (Dwork et al., 2011).

Acknowledgements

The authors thanks Cynthia Dwork for teaching a class which inspired this project, many thoughtful discussions on fairness in generative models, and her suggestions to improve the clarity of this manuscript. The authors would also like to thank Dwork's reading group for providing helpful comments on the manuscript. ML would also like to thank Sitan Chen for many conversations on the theory of diffusion models.

Social Impact Statement

As diffusion models become more broadly used in diverse contexts and applications, it is important that their outputs follow various fairness desiderata. This paper studies several of these desiderata, and provides new methodologies and theoretical frameworks for auditing fairness in diffusion models. We also conduct extensive experiments on real diffusion models. We look forward to future research that will further investigate the questions we raise and consider and provide inclusive solutions.

References

Bakr, E. M., Sun, P., Shen, X., Khan, F. F., Li, L. E., and Elhoseiny, M. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models, 2023.

Berg, H., Hall, S. M., Bhalgat, Y., Yang, W., Kirk, H. R., Shtedritski, A., and Bain, M. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning, 2022.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings, 2016.

Chen, M., Liu, Y., Yi, J., Xu, C., Lai, Q., Wang, H., Ho, T.-Y., and Xu, Q. Evaluating text-to-image generative models: An empirical study on human image synthesis, 2024.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=zyLVMgsZ0U_.

Choi, Y., Park, J., Kim, H., Lee, J., and Park, S. Fair sampling in diffusion models through switching mechanism, 2024.

Chuang, C.-Y., Jampani, V., Li, Y., Torralba, A., and Jegelka, S. Debiasing vision-language models via biased prompts, 2023.

Dehdashtian, S., Wang, L., and Boddeti, V. N. Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss, 2024.

Dev, S. and Phillips, J. Attenuating bias in word vectors, 2019.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness, 2011.

Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., and Kersting, K. Fair diffusion: Instructing text-to-image generation models on fairness, 2023.

Gautam, S., Venkit, P. N., and Ghosh, S. From melting pots to misrepresentations: Exploring harms in generative ai, 2024.

Gokaslan, A., Cooper, A. F., Collins, J., Seguin, L., Jacobson, A., Patel, M., Frankle, J., Stephenson, C., and Kuleshov, V. Commoncanvas: An open diffusion model trained with creative-commons images. *arXiv preprint arXiv:2310.16825*, 2023.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Hodosh, M., Young, P., and Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899, 2013.

Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., and Smith, N. A. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Lee, H., Yoon, S., Dernoncourt, F., Bui, T., and Jung, K. UMIC: An unreferenced metric for image captioning via contrastive learning. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 220–226, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.29. URL https://aclanthology.org/2021.acl-short.29.

Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J. S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H. B., Bellagente, M., Kang, M., Park, T., Leskovec, J., Zhu, J.-Y., Fei-Fei, L., Wu, J., Ermon, S., and Liang, P. Holistic evaluation of text-to-image models, 2023.

Levinboim, T., Thapliyal, A. V., Sharma, P., and Soricut, R. Quality estimation for image captions based on large-scale human evaluations. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3157–3166, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.253. URL https://aclanthology.org/2021.naacl-main.253.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

Li, J., Hu, L., Zhang, J., Zheng, T., Zhang, H., and Wang, D. Fair text-to-image diffusion via fair mapping, 2024.

Luccioni, A. S., Akiki, C., Mitchell, M., and Jernite, Y. Stable bias: Analyzing societal representations in diffusion models, 2023.

Manzini, T., Yao Chong, L., Black, A. W., and Tsvetkov, Y. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1062. URL https://aclanthology.org/N19-1062.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models, 2021.

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., and Marco, F. Bias in word embeddings. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 446–457, 2020.

Perera, M. V. and Patel, V. M. Analyzing bias in diffusion-based face generation models, 2023.

Qiu, H., Dou, Z.-Y., Wang, T., Celikyilmaz, A., and Peng, N. Gender biases in automatic evaluation metrics for image captioning, 2023. URL https://arxiv.org/abs/2305.14711.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

Shen, X., Du, C., Pang, T., Lin, M., Wong, Y., and Kankanhalli, M. Finetuning text-to-image diffusion models for fairness, 2024.

Smith, L. N. and Topin, N. Super-convergence: Very fast training of neural networks using large learning rates, 2018.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Wan, Y., Subramonian, A., Ovalle, A., Lin, Z., Suvarna, A., Chance, C., Bansal, H., Pattichis, R., and Chang, K.-W. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation, 2024.

Wang, J., Liu, Y., and Wang, X. E. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search, 2021.

Wang, J., Liu, X. G., Di, Z., Liu, Y., and Wang, X. E. T2iat: Measuring valence and stereotypical biases in text-to-image generation, 2023.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023.

Yarom, M., Bitton, Y., Changpinyo, S., Aharoni, R., Herzig, J., Lang, O., Ofek, E., and Szpektor, I. What you see is what you read? improving text-image alignment evaluation, 2023.

A. Technical Background on Diffusion Models

In this section, we review the basics of diffusion models and provide background for our theoretical results and experiments.

At a high level, diffusion models seek to generate samples from a distribution given samples from that distribution; e.g. they learn to produce novel images of firefighters given a dataset of images of firefighters. They consist of a forward "noising" process, which iteratively transforms the original distribution by adding small amounts of Gaussian noise to the sample, and a reverse process, which gradually transforms the pure noise distribution into the original distribution through a "denoising" procedure. Intuitively, if one can (e.g. with a neural network) "learn" the noise that is added to images in the initial phase, one should be able to recover an image by starting with pure Gaussian noise and subtracting the learned noise over time (Ho et al., 2020). In our experiments in this paper, we make a standard choice of denoising the images for $T = 1000$ steps. The denoising function at time t is closely related to the statistical *score* $\nabla \ln q_t(x)$ of the noised distribution $q_t(x)$ (Song et al., 2020). Mathematically, the convergence to the true distribution of the sampling process with a sufficiently good approximation of the score can be proved with Girsanov's thoerem (see Section A in the Appendix); empirically, the process is implemented via a discrete-time approximation, where we iteratively denoise images over small time steps. In practice, diffusion models consist of the learned denoising function as well as an image encoder and decoder. The image decoder and encoder maps images to a latent space over which the diffusion process is applied. We also often want to guide the diffusion model towards a particular output; to do this, the diffusion model can be *conditioned* to produce images consistent with a text prompt by allowing the denoiser to depend on a text embedding vector.

Here, we build up to Theorem A.2, which we use as a black box in proving 4.2. Consider a distribution q over $\mathbb{R}^d$ with a smooth density function. Let $\mathcal{N}(0, \mathrm{Id})$ represent the multivariate normal distribution in $\mathbb{R}^d$. Different forward proceses of diffusion models are equivalent to the Ornstein-Uhlenbeck (OU) process up to a reparameterizing of time. The OU process takes samples from q and transforms them into $\mathcal{N}(0, \mathrm{Id})$ by the following stochastic differential equation (SDE),

$$\mathrm{d}X_t = -X_t \,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad X_0 \sim q\,, \tag{1}$$

where the original sample X_0 is taken from q, $(X_t)_{t\geq 0}$ are the time-indexed random variables produced by the SDE after t time, and $(B_t)_{t\geq 0}$ is the standard Brownian motion. We also define q_t to be the distribution of X_t. Intuitively, the OU process dilutes the signal from the original sample X_0 with the $-X_t\,\mathrm{d}t$ term and gradually replaces it with the randomness in $\sqrt{2}\mathrm{d}B_t$. By the forward convergence of the OU process, the KL divergence between q_t and $\mathcal{N}(0, \mathrm{Id})$ decays to 0 exponentially quickly as $t \to \infty$. Now we consider the reverse process of the above SDE, in the sense that the marginal distributions q_t are the same for both SDEs up to smooth test functions. We first fix a terminal time $T > 0$ and consider the following SDE on the time interval $[0, T]$,

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2\nabla \ln q_{T-t}(X_t^{\leftarrow})\}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t\,, X_0^{\leftarrow} \sim q_T. \tag{2}$$

where we start from samples $X_0^{\leftarrow} \sim q_T$ drawn from the noised q_T and $\nabla \ln q_t$ is the score function of the distribution q_t at time t. We again take $(B_t)_{t\geq 0}$ to be the standard Brownian motion. It is well known that q_{T-t} is the distribution of $X_t^{\leftarrow}$, which provides us with a general recipe for sampling from q: we start with samples $X_0^{\leftarrow} \sim q_T$ and then apply the SDE in Eq.2 to generate a sample from q. In practice, one does not have direct access to the score function $\nabla \ln q_t$ of the data distribution and must instead estimate it from training data. This can be estimated by appealing to Tweedie's formula, which relates the score to a denoising problem. We take samples $x \sim q$ and add noise $\eta \sim \mathcal{N}(0, \mathrm{Id})$ to form $e^{-t}x + \sqrt{1 - e^{2t}}\eta$. We train a neural network $\mathrm{NN}_\theta(\cdot, t)$ to estimate x from $e^{-t}x + \sqrt{1 - e^{2t}}\eta$. By Tweedie's formula, the "denoising" function which minimizes the mean squared error of $\mathbb{E}_{x\sim q,\eta\sim\mathcal{N}(0,\mathrm{id})}[\|x - \mathrm{NN}_\theta(\tilde{x}, t)\|^2 | \tilde{x} = e^{-t}x + \sqrt{1 - e^{2t}}\eta]$ can be rearranged to yield the score up to known linear factors.

Lemma A.1 (Tweedie's formula)**.** *Given $\tilde{x} \sim x + e$ for $x \sim p$ and $e \sim \mathcal{N}(0, \sigma^2\mathrm{Id})$, the expectation of $x|\tilde{x}$ is*

$$\mathbb{E}[x|\tilde{x}] = \tilde{x} + \sigma^2 \nabla \ln \tilde{p}(\tilde{x}),$$

where $\tilde{p}$ is the distribution of $\tilde{x}$.

Until (Chen et al., 2023), it was unclear whether this L2-approximation of the true score learned by the neural network could be used to faithfully sample from the base distribution. The below theorem, a simple consequence of Girsanov's theorem, gave an affirmative proof relating distribution learning to adequately solving the denoising problem up to average L2 error.

Theorem A.2 (Section 5.2 of (Chen et al., 2023)). *Let $(Y_t)_{t\in[0,T]}$ and $(Y'_t)_{t\in[0,T]}$ denote the solutions to*

$$\mathrm{d}Y_t = b_t(Y_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad Y_0 \sim q$$
$$\mathrm{d}Y'_t = b'_t(Y'_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}B_t\,, \qquad Y'_0 \sim q\,.$$

Let q and q' denote the laws of Y_T and Y'_T respectively. If b_t, b'_t satisfy that $\int_0^T \mathbb{E}\,\|b_t(Y_t) - b'_t(Y_t)\|^2\,\mathrm{d}t < \infty$, then $\mathrm{KL}(q\|q') \le \int_0^T \mathbb{E}\,\|b_t(Y_t) - b'_t(Y_t)\|^2\,\mathrm{d}t$.

This manuscript will rely on Theorem A.2 as a black-box for future proofs. There are also several more discrepancies between this conceptual model of diffusion models and their practical implementation. First, q_T is unknown and in practice one replaces it with the normal Gaussian. Second, one cannot perfectly simulate the SDE in Eq.2 and must rely on numerical approximations. The errors incurred by both differences can be subsumed into the final distributional error bound with polynomial scaling with respect to the relevant parameters (Chen et al., 2023).

B. Details from Section 4

B.1. Biased Embeddings Correlate with Representationally Imbalanced Generations

To provide empirical support for Theorem 4.2, we study gender bias in occupations for CLIP embeddings used as inputs for SD2.1 and images generated by SD2.1. We consider the `Professions` dataset containing portraits of people in 146 different professions from SD2.1 (Luccioni et al., 2023). For each profession, we compute the ratio of cosine similarity between the CLIP embeddings of the prompts '`Profession, man`' and '`Profession`,' and the cosine similarity of the CLIP embeddings of the prompts '`Profession, woman`' and '`Profession`.' If this ratio is above 1, then this profession is biased in embedding space towards men over woman. To measure the bias in SD2.1 outputs, we sort the image generations into men or women categories with CLIP-ViT-B/32. In Figure 1, there is a mild correlation between gender bias in CLIP embeddings and gender bias in image generations for a given occupation.

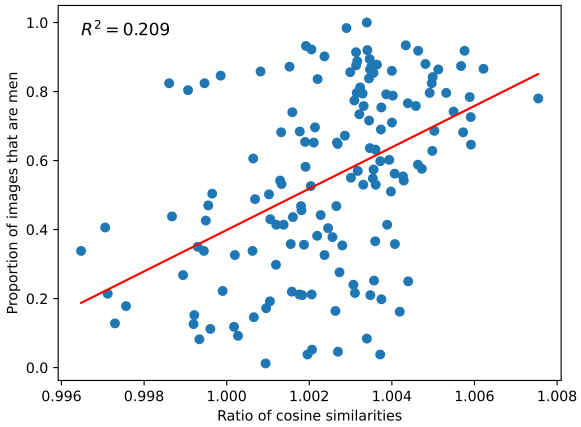

Figure 1. Each point represents a profession from the `Professions` dataset (Luccioni et al., 2023). The x-value is the ratio of cosine similarities between original and gendered versions of the prompt, and the y-value is the proportion of images that are classified as men. Line of best fit is in red and R-squared is reported.

B.2. Additional Details of Diffusion-From-Scratch Training

To better disentangle whether our experimental results above are from a biased embedding or a biased set of training data, we train a conditional diffusion model from scratch with balanced training data across classes but with biased prompt embeddings. Using this model, we test whether generations from this model for a given biased prompt are imbalanced.

To probe this, we first took the the w2vNEWS embedding studied in (Bolukbasi et al., 2016), and in particular train our diffusion model on men and women in three categories: nurses (the second-highest female-biased occupation in (Bolukbasi

Figure 2. For each category in the six classes, we sampled three random images among the 6000 synthetic training images and display them here.

et al., 2016)), philosophers (the fourth-highest male-biased occupation in (Bolukbasi et al., 2016)), and generic people. See Table 2 for the cosine similarities between these embeddings. Within each of these three categories, we trained on 12000 images, 6000 men and 6000 women. Because we couldn't get these images organically, we made a synthetic dataset using Stable Diffusion. We highlight here that for each subject (nurse, philosopher, or generic person), we wanted man vs. woman to be the main difference in image distribution. As such, we controlled all generations to be white, with black hair, and have the same instructions (see below). This is not to say that there are no differences *between* classes; one thing we notice, for instance, is that the nurses appear much younger and the philosophers appear much older. For the purposes of our experiment, however, these differences are not important, as we only label men and women. We also note that we specify the clothing and background for each subject to be clearly recognizable, so that data is easy to label. For instance, a generated image with someone in blue scrubs is clearly a nurse, and a generated image with a library background is clearly a philosopher.

To operationalize our diffusion model, we implemented a conditional diffusion model from scratch based on public reference implementations, re-implemented core modules to handle our biased input embedding, and trained it on the 36000 images. Because of compute limitations, we could only train a model to output 64x64 images. See Section B.2.2 in the Appendix for full details on these changes.

Data labeling was a challenge, because generated model images were small (64x64). Additionally, generation quality was somewhat poor at times (as one might expect with a bespoke model using synthetic data); for instance, some generated image did not contain an actual person. See Figure 3. As such, to label our data as man/woman, we first manually filtered generations for low quality (e.g. pictures where one cannot discern an actual person present), unidentifiable gender, or inconsistent occupation; then, we adopted a consensus-based approach with two reviewers to sort the image generations into men or women groups. For 170 generated images for the following three prompts ("nurse", "person", and "philosopher"), our consensus approach yielded 109 "nurse" images, 130 "philosopher" images, and 139 "person" images.

Table 1 illustrates our results. We see that, even with a balanced set of training data, the majority of nurses were classified as women (59.6%) and the majority of philosophers were classified as men (55.4%). Simply generating an image conditioned on "person," however, yields balanced representation. The frequency of men or women groups for both philosophers and nurses were both equal in the training data, but the diffusion model still exhibits a bias towards female nurses over male nurses and male philosophers over female philosophers. Because we controlled for imbalances in training data distribution in this experiment, this confirms our hypothesis that biased embeddings alone can cause biased outputs.

B.2.1. Synthetic Data Generation.

To get 6000 images across six different classes, we used Stable Diffusion to create synthetic data. In general, we thought these images were high quality and could be used for our training pipeline; Figure 2 illustrates three random images from each of the six classes. Note that it is possible that the synthetic training data itself introduced biases, but the generation process was specifically designed to be as consistent/balanced as possible. Because our model is trained to generate 64x64

Figure 3. We illustrate eight random samples from each of the three classes generated from our trained diffusion model.

Generation	Proportion
Nurse	0.596
Person	0.504
Philosopher	0.446

Table 1. Proportion of women in 200 generations from each class, sampled from our diffusion model trained with a biased text embedding but an unbiased dataset.

images, we then downscaled the output images to 64x64. The precise prompts used to generate 6000 images of each class are listed below:

- **Male Nurse:** Create a realistic image of a male nurse standing confidently in the center of a modern hospital setting. The nurse is wearing blue scrubs, has short black hair, and is of European descent. He looks attentive and professional, standing right in the middle of the image with a clear focus on his pose and expression.
- **Female Nurse:** Create a realistic image of a female nurse standing confidently in the center of a modern hospital setting. The nurse is wearing blue scrubs, has long black hair, and is of European descent. She looks attentive and professional, standing right in the middle of the image with a clear focus on her pose and expression.
- **Male Philosopher:** Create a realistic photo image of a Caucasian male philosopher, situated in the center of a classic library background. He wears a tweed jacket and has short black hair, neatly styled. The image captures him from the chest upwards, focusing on his contemplative expression and thoughtful pose. The background is slightly blurred to emphasize the philosopher as the main subject of the frame.
- **Female Philosopher:** Create a realistic photo image of a Caucasian female philosopher, positioned in the center of a library background. She wears a tweed jacket and has long black hair, neatly styled. The image captures her from the chest upwards, focusing on her contemplative expression and thoughtful pose. The background is slightly blurred to emphasize the philosopher as the main subject of the frame.
- **Man, Generic:** Create a realistic photo image of a Caucasian man wearing a gray shirt, positioned in the center of a neutral background. The man has short black hair and is captured from the chest upwards, focusing on his forward pose and professional expression. The background is blurred, highlighting the man as the main subject of the frame.
- **Woman, Generic:** Create a realistic photo image of a Caucasian woman wearing a gray shirt, positioned in the center of a neutral background. The woman has long black hair and is captured from the chest upwards, focusing on her

	Nurse	Person	Philosopher
`Man`	0.255	0.534	0.290
`Woman`	0.441	0.547	0.176

Table 2. Cosine similarities between man/woman and nurse/person/philosopher in the w2vNEWS embedding.

forward pose and professional expression. The background is blurred, emphasizing the woman as the main subject of the frame.

B.2.2. MODEL CONFIGURATION.

Adding in our embedding. The base implementation of a conditional diffusion model that we base our implementation on takes as input to the UNet (the neural network that learns the score) some vector $v \in \mathbb{R}^{256}$ where $v = f_{\text{pos}}(t) + f_{\text{embed}}(x)$, where $f_{\text{pos}} : \mathbb{R} \to \mathbb{R}^{256}$ is a sinusoidal positional encoding of the time and $f_{\text{embed}} : \mathbb{R} \to \mathbb{R}^{256}$ is an embedding of class (a number from 0 to 9) in 256-dimensional space that gets learned as the model runs. In our diffusion model, we changed the positional encoding of time to instead be of the form $f_{\text{pos}} : \mathbb{R} \to \mathbb{R}^{128}$, and concatenate the positional encoding with a text embedding of the input prompt in $\mathbb{R}^{128}$. Since every embedding in w2vNEWS is in $\mathbb{R}^{300}$, we use a Johnson-Lindenstrauss projection to reduce its dimension to 128.

Training on Multiple Words. We use the w2vNEWS embedding here because (Bolukbasi et al., 2016) conducts a rigorous study of biases in this embedding, and we failed to find such a complete study for CLIP's text embeddings. Because we could only condition on *words* in the embedding, however, mechanically we had to train every training sample/image on three words that labeled the image; see below:

- Male Nurse: Man, Nurse, Person
- Female Nurse: Woman, Nurse, Person
- Male Philosopher: Man, Philosopher, Person
- Female Philosopher: Woman, Philosopher, Person
- Woman (Generic): Woman, Person, Woman
- Man (Generic): Man, Person, Man

Training Details. We used standard hyperparameters found in (Nichol & Dhariwal, 2021) and elsewhere.

- Batch size of 50.
- Max learning rate of $1\text{e} - 4$ on a 1cycle learning rate scheduler (Smith & Topin, 2018).
- Keep an exponential moving average of models for stability, with $\beta = 0.995$.

C. Details from Section 4

C.1. Multicalibration

In Section 5.1, we define a notion of fairness based on the multiaccuracy framework. However, note that the protection offered by multiaccuracy in the example above is fairly weak. In particular, consider the case where $\mathcal{C} = \{I_1, I_2\}$, the true quality of images in I_1 and I_2 is (roughly) uniformly distributed between 0 and 1, and consider the auditing function

$$s(b, i) = \begin{cases} s^*(b, i) & \text{if } i \in I_1 \\ 0.5 & \text{if } i \in I_2 \end{cases}$$

Note that s is $(\mathcal{C}, 0)$-multiaccurate since $\mathbb{E}_{i \sim I_2}[s(b, i)] = \mathbb{E}_{i \sim I_2}[s^*(b, i)] = 0.5$. However, s clearly performs much worse on images of male doctors than female doctors. In particular, assuming that our generative model is reasonably good, we would expect all images i generated for the prompt $b =$ 'doctor' to have a true score $s^*(b, i) > 0.5$, so our auditing function s would consistently give male doctors a higher score than female doctors. Note that this issue can be partially aleviated by defininining a richer class of images $\mathcal{C}$. However, it is also possible to define a stronger notion of fairness that avoids this issue.

Definition C.1 (Multicalibration)**.** Let $\mathcal{C} \subseteq 2^{\mathcal{I}}$ be a collection of subsets of $\mathcal{I}$ and $\alpha \in [0, 1]$. An auditing function s is $(\mathcal{C}, \alpha)$-multicalbirated for prompt $b \in \mathcal{P}$ if, for all $I \in \mathcal{C}$ and for all $I_v = \{i \in I \mid s^*(b, i) = v\}$ where $v \in [0, 1]$,

$$\left| \mathop{\mathbb{E}}_{i \sim I_v} [s^*(b, i) - s(b, i)] \right| \leq \alpha.$$

Note that multicalibration is equivalent to multiaccuracy with $\mathcal{C}$ defined as the level sets of true alignment scores of the original subsets of images. Thus, this definition implicitly creates a richer class of images. Moreover, note that the example function s defined above is not $(\mathcal{C}, \alpha)$-multicalibrated for any $\alpha < 0.5$, since for $v = 1$, $\mathbb{E}_{i \sim I_{2,v}}[s^*(b,i) - s(b,i)] = 0.5$. However, while this is a stronger notion of fairness, there are still ways in which a function that is $(\mathcal{C}, \alpha)$-multicalibrated may behave differently on sets in $\mathcal{C}$; for example, it is possible that a function s that is $(\mathcal{C}, \alpha)$-multicalibrated has more variance in its scores for images of female doctors than male doctors with some fixed true score v.

C.2. Average-then-Score

Here, we define *average-then-score*, an alternative method for calculating alignment scores based on multimodal embeddings. In this approach, to calculate the alignment score of a model M for a prompt b, we start by sampling a set I of images from $M(b)$. We then calculate $\bar{e} = \sum_{i \in I} e_{\mathcal{I}}(i)$. We return the score $\cos(e_{\mathcal{P}}(b), \bar{e})$ scaled to $[0, 1]$. Note that the key difference is when we take the average - in the original approach, the average is taken after the scores are calculated, whereas here we take the average of the image embeddings before calculating the score. To understand why these differ, we show the following result.

Theorem C.2. *For a prompt b and a set of images I, if $\cos(e_{\mathcal{P}}(b), e_{\mathcal{I}}(i)) \geq 0$ for all $i \in I$, average-then-score is lower bounded by score-then-average.*

Proof. We prove a more general result. Given unit vectors $\mathbf{u}$ and $\mathbf{v}_1, ..., \mathbf{v}_n$, where $\cos(\mathbf{u}, \mathbf{v}_i) \geq 0$ for all $i \in [n]$, we show that

$$\frac{1}{n} \sum_{i \in [n]} \cos(\mathbf{u}, \mathbf{v}_i) \leq \cos\left(\mathbf{u}, \sum_{i \in [n]} \mathbf{v}_i\right).$$

First, we see that

$$\begin{aligned} \frac{1}{n} \sum_{i \in [n]} \cos(\mathbf{u}, \mathbf{v}_i) &= \frac{1}{n} \sum_{i \in [n]} \mathbf{u} \cdot \mathbf{v}_i \\ &= \frac{1}{n} \cdot \mathbf{u} \cdot \sum_{i \in [n]} \mathbf{v} \end{aligned}$$

On the other hand, we see that

$$\begin{aligned} \cos\left(\mathbf{u}, \sum_{i \in [n]} \mathbf{v}_i\right) &= \frac{\mathbf{u} \cdot \sum_{i \in [n]} \mathbf{v}_i}{\|\mathbf{u}\| \| \sum_{i \in [n]} \mathbf{v}_i \|} \\ &= \frac{1}{\| \sum_{i \in [n]} \mathbf{v}_i \|} \cdot \mathbf{u} \cdot \sum_{i \in [n]} \mathbf{v}_i \end{aligned}$$

Finally, since $\|\mathbf{v}_i\| = 1$, by triangle inequality we know that $\| \sum_{i \in [n]} \mathbf{v}_i \| \leq n$. Thus,

$$\frac{1}{n} \leq \frac{1}{\| \sum_{i \in [n]} \mathbf{v}_i \|},$$

completing the proof.

To show the original theorem statement, for a prompt b and a subset of images I, consider $\mathbf{u} = e_{\mathcal{P}}(b)$ and $\mathbf{v}_i = e_{\mathcal{I}}(i)$ for all $i \in I$. Scaling the cosine similarity to $[0, 1]$ does not affect the result, so the left side of the inequality is equivalent to score-then-average and the right side of the inequality is equivalent to average-then-score. □

Although this method still biased from the perspective of multiaccuracy, we provide some intuition on why it may reward models that output images with more balanced attributes. Consider a base prompt b and attributes $a_1, ..., a_k$. Let I_ℓ be a subset of images corresponding to prompt b with attribute a_ℓ, and let $I = \bigcup_{\ell \in [k]} I_\ell$. Since the goal of the embedding space is to map prompts close to images that match it, we should expect that $e_{\mathcal{P}}(b)$ is close to $e_{\mathcal{P}}(i)$ for all $i \in I$. Thus, we should also expect that $e_{\mathcal{P}}(b)$ is positioned somewhere 'between' the clusters of vectors corresponding to each subset of images,

though this vector may be closer to some clusters than others. If this intuition is correct, averaging image vectors with more diverse attributes should bring us closer to the vector for the prompt than averaging image vectors with a single attribute, so our score function should reward some amount of diversity, though the exact ratios at which it is maximized may differ. We explore this hypothesis and evaluate this method in Appendix C.3.

One downside of using average-then-score over score-then-average is that it fails to take into account variance. In particular, a set of image vectors could combine to give a very high score because they happen to average in the same direction as the embedding of the prompt even though none are individually close to the prompt. Thus, if used in practice, it will be important to add a term that penalizes variance in the image vectors.

C.3. Evaluation Results

C.3.1. TEXT-TEXT BIAS

As shown in Section 5.2, biases on attributes in the prompt embedding space imply bias for images close to the attributes in the image embedding space. Thus, we start by analyzing the bias present in the prompt embedding space $e_{\mathcal{P}}$ of CLIP. Our results are shown in Table 3.

Occupation	Male	Female	Delta	Average
firefighter	0.971	0.919	0.052	0.959
chemist	0.962	0.923	0.039	0.955
chef	0.954	0.918	0.036	0.950
architect	0.957	0.924	0.033	0.955
biologist	0.978	0.949	0.029	0.972
professor	0.968	0.950	0.018	0.966
doctor	0.962	0.947	0.015	0.965
teacher	0.962	0.947	0.015	0.963
librarian	0.962	0.951	0.011	0.969
hairdresser	0.951	0.958	-0.007	0.967
receptionist	0.954	0.962	-0.008	0.970
nurse	0.951	0.973	-0.022	0.974

Table 3. Text-Text Bias in CLIP

We evaluate 12 occupations for bias on two genders - male and female. For every occupation b, the male column shows the cosine similarity between the embedding of b and the embedding of 'male' $+ b$, and the female column shows the cosine similarity between the embedding of b and the embedding of 'female' $+ b$. The entries are sorted in increasing order of delta, the difference between the male and female similarity scores. We see that our results match several of the biases observed by (Bolukbasi et al., 2016). In particular, 'nurse,' 'receptionist' and 'hairdresser' are *she* professions, while 'doctor' and 'architect' are *he* professions. The average column measures the cosine similarity between the embedding of b and the average of the embeddings of 'male' $+ b$ and 'female' $+ b$. It is interesting to note that these are always closer to the larger similarity score, and are sometimes larger than both. This motivates the average-then-score approach discussed in Appendix C.2 since we would hope that the image embeddings behave similarly.

C.3.2. TEXT-IMAGE BIAS

In this section we investigate the bias in the similarities between occupation prompt vectors and image prompt vectors with specific genders. Our results are shown in Table 4.

Occupation	Male	Female	Delta	Average
doctor	0.800	0.780	0.020	0.801
nurse	0.772	0.833	-0.061	0.813

Table 4. Text-Image Bias in CLIP

We evaluate doctors and nurses for bias on two genders - male and female. To do so, we collect images of male doctors, female doctors, male nurses and female nurses. We then calculate the average cosine similarity between the embedding of

'doctor' and the embeddings of all male and female medical professionals respectively. Next, we do the same for 'nurse'. Note that the set of images is the same across both occupations, so there is no inherent difference in quality; by symmetry, we should expect that images of nurses are as close to the prompt 'doctor' as images of doctors are to the prompt 'nurse.' This is also why we restrict our attention to these two occupations; without the control of using the same sets of images for different professions, it is possible that the male pictures happen to be higher quality than the female pictures of vice-versa.

The delta column shows that for the same set of images, male medical professionals got a 0.02 higher score than female medical professionals on average for the prompt 'doctor,' and a 0.061 lower score for the prompt 'nurse.' Thus, this clearly demonstrates an underlying bias in the auditing CLIPScore function; models that generate only male doctors and only female nurses get over 5% higher alignment scores.

The average column measures the cosine similarity between the occupation o and the average of the embeddings of the male and female images. We see that the average of the images performs better than either individual gender for the prompt 'doctor,' but lands somewhere in between for the prompt nurse. We explore the implications of this further in the next section.

C.3.3. Mitigating Bias

In this section, we explore how our proposed auditing methods, average-then-score and subclass-score, compare to the original method score-then-average on the prompts 'doctor' and 'nurse.' The results are visualized in Figure 4. We find that average-then-score seems to perform roughly as well as score-then-average, biased towards male images for 'doctor' and female images for 'nurse.' However, subclass-score performs significantly better for both, staying roughly consistent as the gender ratio changes. Thus, subclass-score is the most promising step for alleviating gender bias in alignment scores for text-to-image models.

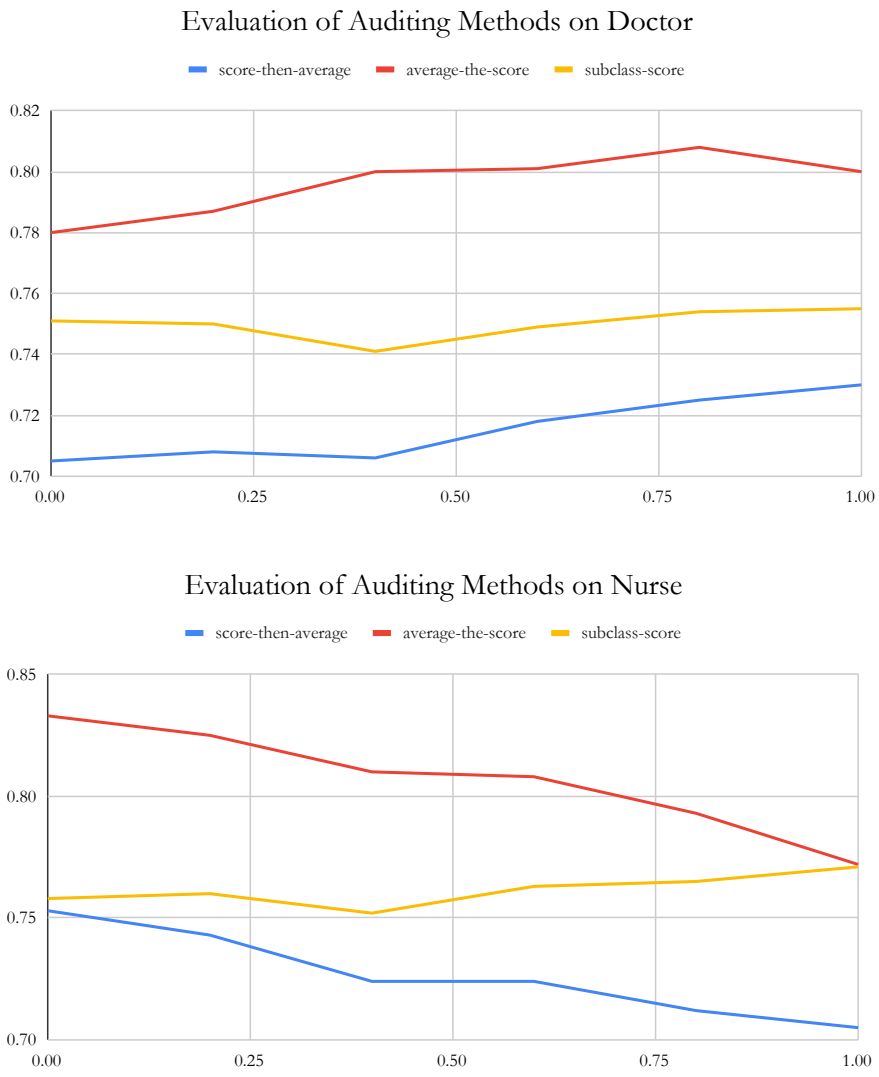

Figure 4. Bias in auditing methods. The x-axis represents the proportion of the images that are male and the y-axis represents the score.