FAIR DECISION MAKING

DAVID YANG†, SELENA SHE†, AMY FENG†, ALEXANDER GOSLIN†

Note: † denotes equal contribution.

1. General Background Reading

1.1. Generative Adversarial Networks.

Definition 1 (Generator)
A **generator** is a neural network that learns to generate realistic samples by transforming random noise into data samples that resemble the training data.

Definition 2 (Discriminator)
A **discriminator** is a neural network that aims to distinguish between real and fake examples.

More formally, a generative model captures the data distribution and a discriminative model estimates the probability that a sample came from the training data rather than G. ”The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency.”

Definition 3 (Adversarial Nets)
Adversarial nets refer to the specific case where the generative model generates samples by passing random noise through a multilayer perceptron, and the discriminative model is also a multilayer perceptron.

1.2. Denoising Diffusion Probabilistic Model / Denoising Diffusion Implicit Model.

Other Relevant Resource(s): DDIM vs DDPM Article

Definition 4 (Diffusion Model)
A **diffusion (probabilistic) model** is a parameterized Markov chain trained using variational inference to produce samples matching the data after finite time.

DDPMs are types of generative models that use probabilistic processes to transform the data from a simple distribution to a more complex target distribution. It iteratively refines the initial random distribution, where in each step it removes the noise from the data and eventually ends up creating a realistic sample of data. In comparison to Generative Adversarial Nets (GANs), GANs may suffer from **real mode collapse** in which the generator

Date: Summer 2023.

produces just a small variety of data that is not as diverse as real-world data. On the other hand, diffusion models will not have this problem but may take longer/tend to be more computationally expensive.

> **Definition 5** (DDIM)
> A **Denoising diffusion implicit model (DDIM)** is a more efficient class of iterative implicit probabilistic models with the same training procedure as DDPMs.

The generative process of a DDPM is defined as the "reverse of a Markovian diffusion process," which approximates the reverse of the forward diffusion process (from data to noise) and is much slower than a GAN, which typically requires only one pass through a network. On the other hand, a DDIM allows for much faster sampling while keeping an equivalent training objective, so that generative models using this architecture are competitive to GANs at the same model size/sample quality." This is done by "estimating the addition of multiple Markov chain jumps by estimating the sum of Gaussian Markov jumps as Gaussian." This generalizes the Markovian forward diffusion process of DDPMs to a non-Markovian one.

One of the key differences between the DDIM and DDPMs is the "**consistency**" property (present in DDIMs but not DDPMs): if we start with the same initial latent variable and generate several samples with Markov chains of various lengths, these samples would have similar high-level features. The consistency of DDIMs allow for meaningful image interpolation.

1.3. Debiasing Methods for Fairer Neural Models in Vision and Language Research: A Survey.

> **Definition 6** (Fairness)
> **Fairness** refers to the "absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics."

> **Definition 7** (Bias)
> **Bias** is any unintended behavior resulting from correlation-based processing that ignores further context not explicit in the data.

Different types of bias include **inductive bias** (unavoidable bias inherent to the learning task), **data bias** (bias of representation/sampling/measurement of data towards a group of subjects) and **intersectional bias** (when an underprivileged group defined by a combination of sensitive attribute and dynamics of individuals are considerably distinct than when considering one sensitive attribute at a time).

We can test the embedding of a given prompt for bias to measure bias in language models. Metrics to measure bias include **Direct Bias (DB)**, where we measure the bias as a projection onto a gender subspace, the **Word Embedding Association Test (WEAT)** which measures bias through permutation of two sets of target words and two sets of attribute words (e.g. engineers/nurses and male/female). The **Sentence Embedding Association**

Test (SEAT) is similar to WEAT but works for contextualized word embeddings.

Definition 8 (Existing Debiasing Categorization)

Debiasing approaches are typically divided into three categories: **preprocessing** (transform the data so that the underlying discrimination is removed), **inprocessing** (modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process), and **postprocessing** (performed after training).

Note: we anticipate that our approaches will be of the postprocessing type.

The survey paper introduces a new taxonomy to categorize debiasing methods:

Definition 9 (New Taxonomy for Debiasing Methods)

A **distributional** strategy modifies the dataset prior to training.

An **inferential** strategy that addresses the problem of fairness based on the model outputs, i.e. that discover and remove social biases without requiring further weight optimization or dataset manipulation

There is also a distinction between methods that focus on optimization via training into two categories:

Definition 10 (1-Step and 2-Step Training)

One-step-training includes fair models generated for a particular task via a single optimization procedure, whereas **Two-step-training** includes models where a new training phase must be performed to fix an existing biased model.

There a few methods to debias one-step-training models:

Definition 11 (Groups of One-Step-Learning Debiasing Methods)

Adversarial methods use adversarial examples to teach the model not to resort to undesired biases.

Causal methods leverage knowledge on causal-effect between protected attributes and outcomes to fix model unfairness (e.g. create new examples during training/generate "counterfactual samples").

Disentanglement methods break down features in latent space to manipulate information independently.

Optimization methods include loss function adaptions, the addition of regularization terms, and other modifications for improving weight optimization.

Discussion — Of the four one-step-learning debiasing methods discussed above, we anticipate looking at **Adversarial** methods and **Disentanglement** methods.

An adversarial approach may work similarly to in the following Adversarial Learning Debiasing Method paper, and a related-to-disentanglement approach is discussed in this Debiasing VLMs via Biased Prompts paper (that being said, this approach may be better categorized as **Vector-Space Manipulation**, which is described later).

Inferential models are distinct from distributional models since they intervene during *inference time* (post-training) to make models fairer.

Definition 12 (Inferential Models)

Prompting prepends or alters the model input with specific triggers that stimulate a bias-free result.

Vector-Space Manipulation manipulates the embedding space to remove undesired biases.

Discussion — We've already seen these two approaches (prompting and vector-space manipulation) in other papers; specifically, the Fair Diffusion (2.1) paper is an example of a prompting approach whereas the Debiasing VLMs paper (2.2) paper takes a vector-space manipulation approach.

2. Focused Readings

2.1. Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness.

> **Approach — Fair Diffusion** is a strategy that attenuates biases after the deployment of generative text-to-image models. It shifts a bias, based on human instructions, in any direction yielding arbitrary proportions (e.g. identity groups), by instructing a pre-trained model on fairness during the deployment stage.

Fair Diffusion is evaluated with Hugging Face's Semantic Guidance pipeline, which allows for changes to image generation to be more easily controlled (and thus enable it to stay closer to the original image composition). To classify the gender of the generated images, they use the pre-trained FairFace classifier.

For example, consider the following pairs of images, generated with stable diffusion and fair diffusion:

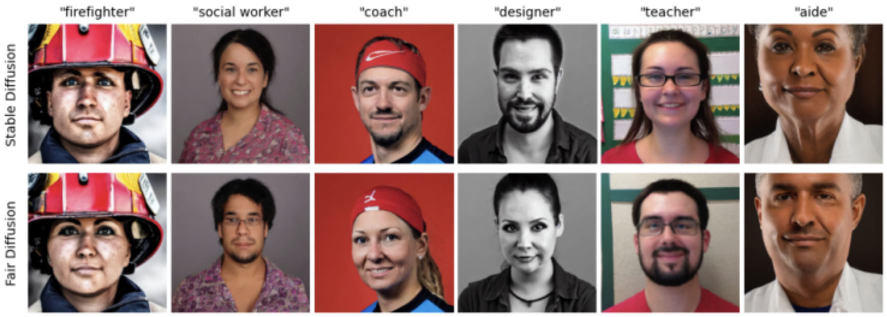
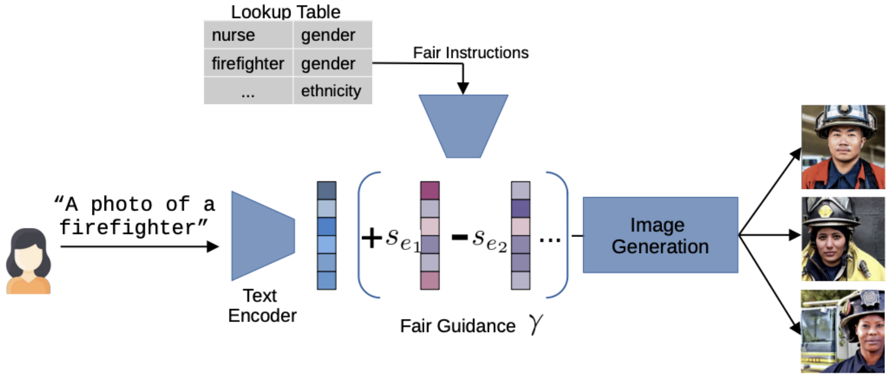


Figure 5: Generated Images with SD (top row) and Fair Diffusion (bottom row) for different occupations. The images are generated with the prompt "A photo of the face of a $\{occ\}$", in which each column name represents the used occupation (occ). For generated images of female-appearing persons, we applied fair guidance with - "female person" + "male person" and vice versa for male-appearing persons. One can observe that Fair Diffusion changes the typical gender appearance for each occupation image while keeping the residual (occupation-related) features present.

We notice that, despite correcting for the protected attribute of gender, the remainder of the generated images' characteristics stay the same.



Lookup Table

nurse	gender
firefighter	gender
...	ethnicity

Figure 1: Fair Diffusion deployment. A user inserts a prompt to generate an image. With the help of fair guidance, the image generation is steered toward a fairer outcome. Here, the fair instructions are realized with a lookup table —the biased concept is recognized, and guidance γ is applied. Like in Eq. 5, the fair instructions, e_i, are transformed into vectors, c_{e_i}, by the text encoder and can be scaled by s_{e_i} to perform fair guidance. Here two editing prompts (purple-colored vectors) are illustrated, but different numbers are possible too. The lookup table can be set up by any user. (Best viewed in color)

As discussed in the paper, fair diffusion works at the following high-level manner: the generated image x satisfies

$$x = \eta(p, \gamma(e, s_e)).$$

More specifically, η is a function of the text input prompt p (e.g., "A photo of a firefighter") with "fair guidance" , γ, given during the generation. In turn, γ depends on additional textual descriptions of attribute expressions e_i, scaled by s_{e_i} with guidance direction. As a result, the image generation is guided towards the input prompt p and fairness instructions e_i simultaneously.

To realize the different expressions of an attribute with Fair Diffusion, the authors control the guidance direction by randomly sampling from a desired probability distribution P. That means each e_i is either increased or decreased depending on the expression that should be promoted/suppressed.

The above figure displays a binary case where concept e_1 is promoted (+) and e_2 is suppressed (-) during the image generation – their direction can be changed based on P. The purpose of the **lookup table** is to identify prompts requiring fair guidance and to align the output with the users' fairness notions.

Another area this paper investigates is "bias inspection" in the components of the well-known text-to-image generation model Stable Diffusion. Stable Diffusion relies on the large-scale image dataset LAION-5B and the pre-trained text encoder model CLIP.

In the subset of the image dataset LAION-5B that the paper uses, they found that the "Science" and "Engineering" subsets have lower rates of female-appearing persons (around 0.35 and 0.2, respectively), while "Arts" and "Caregiving" (around 0.55 and 0.75, respectively) have higher rates of female-appearing persons, reflecting stereotypical gender occupation biases.

Similar biases are revealed after the authors use the iEAT Bias Test for CLIP:

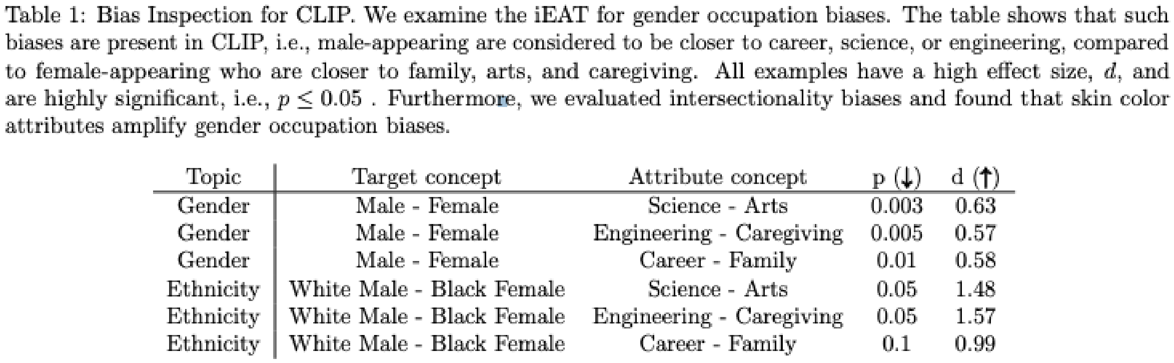


Table 1: Bias Inspection for CLIP. We examine the iEAT for gender occupation biases. The table shows that such biases are present in CLIP, i.e., male-appearing are considered to be closer to career, science, or engineering, compared to female-appearing who are closer to family, arts, and caregiving. All examples have a high effect size, d, and are highly significant, i.e., $p \leq 0.05$. Furthermore, we evaluated intersectionality biases and found that skin color attributes amplify gender occupation biases.

Topic	Target concept	Attribute concept	p (↓)	d (↑)
Gender	Male - Female	Science - Arts	0.003	0.63
Gender	Male - Female	Engineering - Caregiving	0.005	0.57
Gender	Male - Female	Career - Family	0.01	0.58
Ethnicity	White Male - Black Female	Science - Arts	0.05	1.48
Ethnicity	White Male - Black Female	Engineering - Caregiving	0.05	1.57
Ethnicity	White Male - Black Female	Career - Family	0.1	0.99

In fact, bias amplification occurs when the association is modified by ethnic attributes (males as European and females as African-American).

To summarize the results, the authors "found biases and unfairness in each component of the SD pipeline: in the LAION-5B dataset, in the CLIP encoder, and in the generated images. At the same time, the biases are not simply mirrored between LAION-5B and SD's outcome and do not show a clear tendency."

Discussion — The prompts in this paper are all of the format "give me a photo of the face of a [ ]."

A specific ambiguous case mentioned is the "face of a dishwasher" (which gives a photo of a dishwasher rather than someone washing the dishes). Consequently, can the given prompt be modified to "the face of a person washing dishes" (or other similar formats), and how might these shifts affect the paper's results?

Extension — Implicit associations are not accounted for (this paper requires a lookup table mapping pre-defined occupations to protected groups). However, a prompt such as "a photo of a hospital" may generate images where doctors are predominately male.

Can we extend or modify the existing work to support biases implied from text prompts?

2.2. Debiasing Vision-Language Models via Biased Prompts.

Approach — Many approaches for debiasing Vision-Language Models require training or fine-tuning models using resampled datasets, or modified objectives, which can be computationally expensive. The paper proposes a general approach for ”self-debiasing foundation vision-language models by projecting out biased directions in the text embedding.”

This approach creates a projection matrix (where each row is an attribute) that projects out biased directions. To address possible unstabilities caused by prompts defining the biased directions, they use “positive pairs” (pairs of prompts that are expected to have the same semantic meaning after projection) to calibrate the projection matrix. Their overall approach does not require training, data, or labels, so it is computationally efficient.

The bias of a classifier will be quantified by computing the cosine similarity between its weights and the corresponding spurious feature. We expect rows of classifier weights to have similar cosine similarities to pairs of embeddings: for example, the embedding of “a photo of a doctor” should be similar to “a photo of a male doctor” and “a photo of a female doctor.”

Approach — To evaluate generative models, the paper uses the CLIP classifier to predict the sensitive attributes and human evaluation to corroborate these results.

They mention that as an alternative, the FairFace classifier could be used. However, they note that the “domain shift between the FairFace dataset and the generated images significantly impaired performance” and thus used the CLIP classifier.

Discussion — This paper mentions that FairFace dataset and the generated images significantly impacts the bias/general performance (so they used CLIP) but the Diffusion Paper (2.1) used FairFace as the final classifier.

We will need to be cognizant of our choice, specifically how the classifiers work on our generated images.

2.3. **Learning Transferable Visual Models from Natural Language Supervision (CLIP).**

Approach — **CLIP** (Contrastive Language-Image Pre-training), is an efficient method of learning from natural language supervision.

Given a batch of N (image, text) pairs, CLIP is trained to predict which of the $N \times N$ possible (image, text) pairings across a batch occurred. They do this by jointly training an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings of the N real pairs in the batch.

2.4. **Diffusion Models Already Have a Semantic Latent Space.**

Approach — Diffusion models lack a semantic latent space which is essential for controlling the generative process. The authors introduce a asymmetric reverse process (Asyrp) which discovers the semantic latent space in frozen pretrained diffusion models. The semantic latent space, named h-space, has nice properties to accommodate semantic image manipulation: homogeneity, linearity, robustness, and consistency across timesteps.

2.5. **DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation.**

Approach — **DiffusionCLIP** performs text-driven image manipulation using diffusion models. Based on full inversion capability and high-quality image generation power of recent diffusion models, the method performs zeroshot image manipulation successfully even between unseen domains. The authors propose a novel noise combination method that allows straightforward multi-attribute manipulation. Code is available in their DiffusionCLIP GitHub

Other “state-of-the-art text-driven manipulation methods” include TediGAN, StyleCLIP, and StyleGAN-NADA (which leverages GAN inversion and CLIP loss to achieve image manipulation that does not require additional training with simple text prompts).

2.6. **Translating from Unfair to Fair Embeddings.**

Approach — Use a translator network (a neural network) to produce an alternative embedding for a given photo. The translator network assumes nothing about the structure or training procedure for the original embeddings.

The overall goal is to preserve as much of the original embedding as possible while removing any encoding of protected attributes.

Discussion — In our own experiments, we can think about their loss function choice (backed up by other work in literature) and their overall procedure.

3. Summary of Work So Far

3.1. **Face Detectors.** Existing Facial Detection Algorithms are evaluated on the widely known WIDER FACE (Face Detection Benchmark Dataset). A number of Facial Detection Algorithms and their evaluations on WIDER FACE can be found here. Note that some of these algorithms are object detection algorithms adapted for facial recognition, but some are also specifically facial detectors (e.g. RetinaFace, YOLO5Face).