MONASH University

Bias Modelling and Mitigation in Diffusion Models

Sai Kumar Murali Krishnan

Student ID: 30377390

Bachelor of Applied Data Science (Honours)

October 29, 2023

Final Thesis submission for ADS4100 under the supervision of Bhautik Joshi and Yuan-Fang Li

Faculty of Science, Monash University

Contents

Copyright notice

Abstract

Text-to-image models exhibit biases, particularly in occupations, through social attributes such as gender, race, age, or skin tone. Latent diffusion models, known for their high-fidelity image generation, reproduce and amplify these societal biases, impairing one's view of reality and exacerbating harmful stereotypes. We present a novel post-processing mitigation method that modifies word representations at the text encoder of a latent diffusion model. This method reduces biased representations at the image generation stage, favouring fair and more equitable images aligned with a specified fairness policy.

Acknowledgements

This thesis would not have been possible without the assistance of Dr Simon Clarke, who assisted in creating the honours thesis as a final assessment and actively gave feedback on my academic writing throughout the year. I am also deeply indebted to Dr Bhautik Joshi, who was my host during my internship at Canva and then co-supervisor during my thesis. His guidance inspired me to experiment with image generation techniques and later enabled me to continue my research in the form of an honours thesis. A major thanks to Dr Yuan-Fang Li for providing support as a supervisor through suggestions and approaches to formalising my work. I also wish to acknowledge the efforts of Dr Alexis Gray, whose expertise in sociology has given me more perspective on sociology and how it affects artificial intelligence.

I'd also like to thank my team at Canva, the AI Images team, who worked with me extensively to deploy my method to Canva's text-to-image product and as a source of ideas for me to experiment with. I am immensely grateful for the opportunity to intern at Canva.

I want to thank my parents and younger sister for caring for me during my thesis work and providing emotional support to finish my research. Lastly, I acknowledge the many friends who supported me throughout the creation of this thesis this year. Special mentions to Jim Ye and Len Luong for providing me with directions on how to structure my thesis. Special mention to Vicky Mach, Sung Ho Chu, Rita Truong, Tsui Shin Mok, See Jian Shin, Ashley Monaghan, Thi Tran, Chris Dalas Christodoulides, Michelle Hong, James Lew, Zakir Oulad Hadj Tolentino, Rohan Rajesh Kalanje, Rahul Pejathaya, Ali Hasan, James Roche, Madison Geeson, Puti Dai, Junru Wei, Lauren Yim, Indra Kusumah-kasim, and Joshua Broeksteeg for providing moral support and listening to me talk about this topic.

Chapter 1

Introduction

In recent years, image generation models have taken off, producing photorealistic images with a given description, commonly known as a prompt, such as Figure 1.1. In particular, diffusion-based models have broken records for generating photorealistic images partly due to the power of modern transformer-based language models. Diffusion models have demonstrated actual use cases, with various uses beyond image generation, such as 3D modelling, video generation, animation or image editing techniques. As these families of models concerning visual modalities rise to mainstream uses, such as in the design industry through Adobe's Firefly, a generative AI art application (Wiggers, 2023), or more consumer-oriented AI avatar apps, via the rapid deployment of so-called foundation models, issues concerning representation and bias in image generations have begun to emerge, specifically regarding social attributes such as gender, ethnicity, and common stereotypes that embed themselves in a visual medium.

We introduce a method for mitigating biases in the diffusion-based, text-to-image models at the text layer, which foregoes traditional fine-tuning and enables alignment with a given policy, such as encouraging diversity within occupation-based generations. This project proposes systematically debiasing occupations concerning race and gender, then explores the resulting image generations, noting any degradation to image quality by a modelling process.

Modelling biases in image models is accomplished by generating images over a large set of prompts, as seen in Luccioni et al. (2023) or Seshadri et al. (2023), and measuring the disparity between reported job statistics, indicating that models amplify biases. Methods for mitigating bias and working towards algorithmic fairness in generative AI are less developed, particularly concerning the modality

Figure 1.1: *AI photo with the prompt: "Photo of a capybara in a jumpsuit", in Stable Diffusion 2.1*

of images. In contrast to discriminative machine learning, where there is a discernable ground truth that a model should tend towards, generative models attempt to estimate the distribution of the data it was trained on, and aiming for a more diverse representation is a goal that depends on the observer. We use a multimodal image transformer for visual question answering, which integrates the facilities of text understanding from a large language model with image understanding capabilities for querying regions of an image.

1.1 Sociological Insights Into Social Bias in Image Models

While bias mitigation in machine learning is a well-studied topic, social bias in generative models has much less literature on mitigation and modelling. With the progress seen through the large-scale deployment of machine learning systems, societal biases are made more prominent in the systems, which can be legally and morally ambiguous. To begin categorising and distinguishing these biases, we use sociology to rigorously determine what social biases may come from and what they represent.

We briefly examine sociology to explore the dynamics of bias in image models. The objective is to explore gender, ethnicity, and race, among other social attributes. As these attributes are deeply ingrained in our society, they are often used in describing people, a standard training source for image generations where text captions are paired with images of people. However, it is essential to acknowledge that race and ethnicity, which are used interchangeably, represent distinct concepts of a person and their identity.

Race typically refers to a categorisation based on physical characteristics, such as skin tone, hair type, and facial features. The concept arises in many societies as a method for classifying people. However, it is worth noting that race is a social construct that reflects people's classification based on significant physical differences. On the other hand, ethnicity is a broader concept encompassing tradition, language, religion and other non-physical factors. The term refers to a community of people who share these factors, and identifying with these factors is an indicator of belonging to that group. Unlike race, ethnicity is an expression of the person's identity and may not immediately be discernable to other people, let alone a machine learning model.

The challenge of interrogating biases concerning social attributes is that machine learning models often oversimplify and conflate these traits at the data collection or training stage. There is a distinction between "race" and "ethnicity" in sociology, though a model cannot discern between the two traits and often uses one instead of another. A model also discretises these constructs into categories that lose nuance and representation. Much like the biases of a machine learning model, this oversimplification mirrors society, indicating that we must make clear what these attributes are defined as while also exploring how a machine learning model would understand these attributes.

In efforts to increase the diversity of image generation in diffusion models, it is also essential to bridge the gap between machine learning and sociology to understand better the disparity in understanding social attributes such as race and ethnicity. This way, our terminology is accurate, and we have an insight into how AI interchanges and destroys nuance in these terms.

1.2 Prior Work on Mitigation and Modelling

From November 2022 to February 2023, we undertook a project was undertaken on safety and bias mitigation on a text-to-image service. The nature of the internship leaned towards experimental research on diffusion-based models, specifically Stable Diffusion. The internship began with testing

on Safe Latent Diffusion, aimed at mitigating the potential for generating explicit or harmful content. However, in attempting to mitigate harmful content, the quality of images had noticeably degraded.

We then chose to pivot the concept of 'textual inversion'. The paper "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion" (Gal et al., 2023) allows users to personalise Stable Diffusion by providing a small image set, which would then create a new style or concept in the model. We used Stable Diffusion to generate a more representative dataset for a given occupation and then created a more representative embedding using textual inversion. While generations involving the chosen occupations were balanced with gender and skin tone, particular professions were more biased and required hyperparameter tweaking, such as changing the number of steps or modifying the learning rate. The method would be used in the most prompted occupations to make image generations broadly more representative and make the text-to-image service safer to use.

As the internship concluded, we discovered that concept embedding arithmetic was a sensible operation and that there was merit to applying the arithmetic for bias mitigation. The mitigation method bypasses the training process, which takes hours to train a single embedding, and directly operates on the embeddings. This thesis will begin with the internship project and explore more efficient bias mitigation strategies. This thesis also aims to highlight discourse, what it means to mitigate bias in an image model, and the limitations of the proposed method.

1.3 Rationale

As foundational models are trained and then packaged for use, it falls on associations using the models to study and mitigate any potential bias issues for safety reasons. However, the research also indicates that the models are limited by their knowledge and that mitigation is a bandaid for immediate harm but is limited by its discretised understanding of sensitive traits. This thesis aims to propose measures for alleviating unequal representations in generations. Additionally, this thesis aims to explore the understanding of text-to-image systems concerning sensitive traits. There should be discussions around data collection; for example, more work should be afforded to captioning images used to train image generation models.

1.4 Thesis Statement

Diffusion-based image generation models present bias amplifications in various terms concerning social attributes such as gender, race or skin tone. We propose quantitative and qualitative methods of testing and quantifying the extent of bias, specifically in job occupations. Using this framework for measuring bias, we devise a bias mitigation model that increases the diversity of subjects in images concerning given biases.

1.4.1 Aims

Various industries use text-to-image products or similar diffusion-based pipelines with a text-encoder, design applications, entertainment and animation teams, or deployment for user-oriented applications as part of social media platforms. This thesis intends to demonstrate the effectiveness of a training-free process for bias mitigation regarding text embeddings' speed and natural interpretability.

The importance of a training-free method is also due to the training time and computational resources required for foundation models, which are often expensive and inaccessible, typically amounting to months in training time with millions of dollars in computing resources. Re-training a large model with a filtered dataset is not feasible nor accessible to many organisations wishing to safeguard their products. For similar reasons, fine-tuning text-to-image models requires GPU time and much longer to refine. A training-free method can also be used for quicker iteration in mitigating concerns in widely deployed applications.

The thesis does not present an exhaustive solution for biases or harmful representations in images and, as such, asks for more investigation into more rigorously defined attributes. In writing on this topic, we hope that more structured frameworks for examining bias from a well-defined forensic perspective can take off and allow for more specific definitions of biases beyond what comes reported with a model, typically that a model exaggerates biases.

We pose the following research questions:

1. How does manipulating text embeddings affect the resulting images?
2. What terms can be used to balance the representation of images?
3. How does embedding-based mitigation vary from other bias mitigation methods?

1.5 Thesis Structure

The structure of this thesis will encompass several components. It will begin with a comprehensive literature review highlighting the progress in generative artificial intelligence and the expanding body of work concerning bias mitigation in foundational models. Additionally, it will explore historical approaches to fairness in machine learning. The thesis will then feature a chapter on modelling biases in text-to-image models, delving into a methodology for querying image generation models, particularly in identifying social biases. We follow with a significant component of the thesis: an in-depth exploration of an optimisation-based mitigation method designed to enhance the diversity of image generations in the context of occupations. This method will be thoroughly analysed with qualitative and quantitative results, providing valuable insights into the effectiveness of the proposed method.

Chapter 2

Literature Review

Examining current and prevailing literature on mitigating bias in large generative models demonstrates the gap in forensically examining biases in generative AI and new opportunities to explore similar family models. We briefly summarise current results and underline existing issues with various machine learning systems, specifically to indicate the space for development with large-scale image generation models. We then explore the existing literature surrounding mitigation methods for diffusion methods, outline any shortcomings of the proposed methods, and outline where the methods relate to our proposed method.

2.1 Generative Artificial Intelligence and Foundational Models

Gozalo-Brizuela & Garrido-Merchán (2023) examines the development and surge in large-scale generative AI applications through "A survey of Generative AI Applications". Compared to discriminative AI, which typically classifies or determines outcomes based on existing data, generative AI synthesises new data by estimating the underlying distribution of the data. The recent development in computing power via graphics processing units (GPUs) and tensor processing units (TPUs), more powerful architectures such as the transformer or the diffusion model, and a wealth of data has led to so-called foundational models that can take in a variety of inputs such as text, image, video, sound, or custom formats. The survey determines 15 broad categories for the models while noting that many of the models reduce to a text model in some form, assuming that many distinct modalities such as code, business, medical advice, or marketing can be represented as text. The primary categories of interest are visual mediums, primarily the image modality, and, by a faint extension, video models. The space

for generative image models includes explicit image generation. These image editing techniques extend to software such as Adobe Photoshop and various styles in generating images based on an input prompt. The paper also outlines several other modalities, such as 3D models and video, that use Diffusion models for generation, in addition to image generation models. As a separate category, multimodal models combine various modalities, which arise from combining model architectures, such as the Vision Transformer proposed in Dosovitskiy et al. (2021), where a model can query images with text prompts. The paper concludes that there is a revolutionary change in how businesses integrate software, specifically with near-human capabilities, but also emphasises the ethical concerns of using models for sensitive services, such as medical advice.

2.2 Image Generation Networks

The primary subfield examined will include models that generate images and fall under the area of computer vision. We include distinctions between Generative Adversarial Networks (GANs) and Diffusion-based models and the change in paradigm regarding how an image is synthesised. Zhang et al. (2023) examines the current state-of-the-art image models with "Text-to-image Diffusion Models in Generative AI: A Survey". A description of each model, with their respective architectures, is detailed after this section. While there is a focus on Diffusion Models, the history before them consists of GANs and autoregressive methods such as DALL-E. The paper also notes a newer method called classifier-free guidance, where an extra classifier guides the diffusion process and drastically improves the quality of image generations in various pipelines.

Diffusion models can be split based on whether they denoise in the pixel space or incorporate an auto-encoder/decoder and work in the latent space. Across the board, all of the mentioned diffusion boards incorporate text encoders for language understanding and act to condition image generation with text, and as the size of the text encoder model increases, so does the fidelity of the generated image. The size of the text encoder is highlighted to have a significant role as Imagen's text encoder, T5 XXL, which has around 11 billion parameters,compared to CLIP which has 63 million parameters, a text-encoder/image-encoder pair which were trained on a smaller set of image-text data in Radford et al. (2021). As an overarching consensus, the survey indicates that the Frechet Inception Distance (FID) was most commonly used and is a metric measuring the distance between generated and real images. Other benchmarks include human viewing and rating and can test a model's inherent social biases or visual capabilities. Overall, the survey indicates a variety of use cases for diffusion-based

models that extend beyond simple image generation but also notes that there should be a more unified framework for metric evaluation.

2.2.1 Generative Adversarial Network

Goodfellow et al. (2014) describes a method for efficiently training a generative model by setting up an adversarial game between a generator model G and a discriminator model D. The discriminator D determines whether an image is "synthetic" or real, where a synthetic image is generated from the generator model G. The game works by training the discriminator to detect synthetic images better and training the generator to fool the discriminator better. In particular, the generator begins with a random seed consisting of noise and then maps said noise to the space of the dataset. The adversarial game is formulated by minimising the chance of being detected for the generator while maximising the detection chance for the discriminator. With the correct initial conditions, the paper indicates that the learned synthetic distribution should be indistinguishable from the real dataset to the discriminator and thus stabilise. However, the paper notes that models can collapse into a "Helvetica scenario" where a poor training regime causes the generator not to be able to generalise the images it generates. Generally, generative adversarial networks provide a method for drawing random instances from an estimated distribution; however, they are difficult to train effectively. Despite training difficulties, GANs solidified the space for image generation networks.

2.3 Diffusion Models

The standard pipeline for the text-to-image Diffusion model consists of 3 individual models linked to each other, and we discuss each model with their history. Sohl-Dickstein et al. (2015) outlines a method for unsupervised deep learning that takes inspiration from non-equilibrium statistical physics. The concept involves adding noise to some initial entity and then using the neural network to restore the entity via denoising. Adding noise is typically represented as a Markov Chain with several steps T. Training the model comes to finding a method of reversing the transitions of the Chain via gradient descent and consequently attempting to rever the data as faithfully as possible. Ho et al. (2020) builds on the concept of denoising with the paper titled "Denoising Diffusion Probabilistic Models. The proposed training objective is the expected difference between the actual noise and the noise the model decides to remove. As a benchmark, the paper uses the Frechet inception distance to compare generated images to real images, with promising results both to the human eye and computed metrics. It reports higher scores than nearly all image generation models as of 2020, corresponding to the

sampled images' seemingly realistic and precise image quality. The paper closes by indicating that the diffusion models admit a form of data compression, a topic that will be reasoned about in subsequent papers.

Saharia et al. (2022) develops pixel-space diffusion with Google's Imagen in "Photorealistic Text-to-image Diffusion Models with Deep Language Understanding". The paper emphasises the importance of language understanding as the text encoder of choice is the T5-XXL, an 11 billion parameter language model explored in Raffel et al. (2020). Imagen consists of a U-net for generating a 64x64 image, then upsizes the image twice to a 1024x1024 image and modifies the text conditioning at the upscaling networks to use only cross-attention instead of self-attention and cross-attention. The paper also introduces DrawBench as a benchmark and seeks to interrogate social biases in addition to visual reasoning abilities. The benchmark consists of prompts that push the model with esoteric words and long descriptions and observe how the final image generation lines up with the prompt. The paper also trials different variants of T5, with the largest encoder being more critical than the U-Net for denoising. The paper also highlights that T5-XXL was preferred over CLIP by human evaluators, indicating that the text encoder and how text is conditioned into generation are crucial.

The choice of a denoising model primarily impacts the speed of image generation. Saharia et al. (2022) concludes by indicating that the study of language models in the space of image generation has a significant impact on the final generation. The model is not made available for public use as one of its data sources, LAION-400M, contains a range of inappropriate content considered sexually explicit or harmful. Many diffusion models in this review pull on LAION as a data source, a common thread that ties all the models and, consequently, all their bias issues.

2.3.1 Variational Autoencoders

Kingma & Welling (2013) provides the foundation for a non-deterministic compression model in "Auto-encoding variational Bayes". An encoder/decoder paradigm is explicitly used to compress an input by mapping it to a smaller, latent space and decoding the latent image into a reconstruction of the input. The network learns to reconstruct the input while being trained to avoid memorising the input. The latent space is typically modelled as a multivariate distribution, with input data mapped to points in this space. Then, a decoder is trained to revert this mapping or approximate the reversion. An aspect of training the decoder to reconstruct an input is that sampling random noise and then feeding it to a decoder would generate new data based on the distribution it has learned. This

continuous mapping provides a probabilistic interpretation of the latent space instead of deterministic autoencoders locked to a fixed representation. Another related use of the latent space was that one could interpolate between two points in latent space and derive a sensible meaning from the output, which would be a smooth interpolation between two objects, a desirable quality for tasks such as image editing. Overall, the paper is a cornerstone publication in the generative space as they are used in various image generation models such as GANs and Latent Diffusion Models.

2.3.2 Latent Diffusion Models

Rombach et al. (2022) develops on the concept of diffusion models, with the perspective that denoising on the pixel space directly is an inefficient and computationally intense process, and proposes the use of variational autoencoders, which follows from the concept proposed in Kingma & Welling (2013). (Rombach et al., 2022) finds that performing diffusion over the latent space is cheaper, more GPU efficient, and retains their quality. Examining diffusion over pixel space reveals a perceptual compression consisting of high-quality details, followed by a semantic compression, which focuses on capturing concepts and the composition of images. The paper proposes using an autoencoder as a computationally cheaper method for perceptual compression. These models are called Latent Diffusion Models (LDMs) and are easier to scale due to decoupling the two different compression phases. Decoupling the autoencoder from the diffusion model makes using different autoencoders with different diffusion models possible. LDMs are considered an evolution of Diffusion Models and variational autoencoders, which aim to better compress high-quality details and lower computational costs by operating in a lower-dimension latent space. The compression is typically done by a factor f, with a slight compression factor having no effect and a significant compression factor losing too many details. The experiments regarding conditional Latent Diffusion using a text encoder with classifier-free guidance also demonstrate noticeable quality increases, speaking to the effectiveness of using autoencoders as an intermediate representation. (Rombach et al., 2022) closes by stating that LDMs use less computational resources for higher quality images and respond well to cross-attention conditioning for a range of image generation tasks, namely text-to-image generation.

2.4 The Transformer

Vaswani et al. (2017) is a cornerstone work of modern language models, commonly known as foundation models, characterised by their dense understanding of the corpora on which they are trained. As this project examines the intersection of language and vision with both Stable Diffusion

and BLIP using the transformer architecture, the significance of this paper is noted. Vaswani et al. (2017) proposes the attention mechanism for language models, specifically on machine translation. The underlying architecture uses attention heads that reveal relationships between sequence terms. The paper proposes a multi-headed attention mechanism and notes a distinction over recurrent neural networks, which cannot be parallelised due to being sequential. Overall, the works from "Attention is All You Need" laid a foundation for modern foundation models that use attention mechanisms for language tasks. Devlin et al. (2018) develops the applications and structure of the transformer with "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", the same text encoder used in (Rombach et al., 2022). BERT improves on previous models that are unidirectional by training using a "masked language model" pre-training objective. The process involves hiding a random selection of tokens and trains the model to recover the masked tokens. The pre-training task increases the effectiveness of language, such as determining the validity of sentences with each other, checking if sentences are grammatically correct, or checking if sentences are similar. (Devlin et al., 2018) primarily focuses on transfer learning with language models for language understanding, which would later become relevant for Diffusion Models.

2.4.1 Vision Transformers

The paper "An Image Is Worth 16x16 Words: Transformers For Image Recognition at Scale" introduces the Vision Transformer (ViT), combining the transformer architecture with the image modality. Dosovitskiy et al. (2021) forgoes using a convolutional neural network (CNN) to show that processing image patches with a transformer can work just as well for image recognition tasks. The convolutional architecture for visual feature extraction is standard, though the paper replaces convolutions with attention heads. Instead of applying the attention mechanism to each pixel, Dosovitskiy et al. (2021) splits an image into 16x16 patches and feeds the patches into a linear projection into a transformer. The paper replicates Vaswani et al. (2017), except instead, feeding in a sequence of patches.

Compared to a convolutional architecture, there is less inductive bias learned, meaning there is less importance on local features, and all these features are pooled at the attention stage. The vision transformers have up to 632 million trainable parameters and are compared against the ResNet, a staple CNN. The results from vision transformers beat out CNNs while being smaller in terms of weights, often using up to four times less computing power for similar performance. Vision transformers demonstrate excellent ability with few-shot training on various natural language processing tasks.

The paper finds they have the potential for computer vision tasks, which Li et al. (2022) expands on with BLIP and Radford et al. (2021) for jointly training image and text representations.

Li et al. (2022) develops on the vision transformer with "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". BLIP is a vision model that can perform high-quality image captioning and visual question-answering. The underlying architecture builds on the Vision Transformer proposed by Dosovitskiy et al. (2021) and extends functionality with a mixture of encoders and decoders for encoding images and text. An Image-Text Contrastive Loss (ITC) is used to unify vision and language understanding. Then, an image-grounded text encoder takes in a caption of the image and encodes the information. The encoder is paired with an image-grounded text decoder, which forms the basis for text generation. Each encoder uses a different loss function and is united for multi-task learning. The paper achieves state-of-the-art results by sharing all parameters for encoders and decoders, except for the self-attention parameters. Li et al. (2022) introduces CapFilt, a caption filtering system for generating and filtering captions for image-text pairs due to a large amount of noise and low-quality captions in the dataset, which can degrade the performance of BLIP. CapFilt is trained using an image and text transformer and generates synthetic captions that are more diverse in the dataset. Parameter sharing is also employed similarly for CapFilt, with results finding that having different self-attention layers leads to better performance. For BLIP, the tests consist of captioning against other models, though they also venture into visual question answering (VQA). Across the board, BLIP was found to have performance mirroring other state-of-the-art models while lowering the number of trainable parameters and the amount of human annotation required for a clean dataset.

The cross-attention mechanism is also heavily used in diffusion models; the ability to condition an image network with cross-attention has proved very effective. The cross-attention augmentation usually starts with a text encoder, which maps text to an intermediate representation for an image model such as a UNet, as Rombach et al. (2022) does. Innovations in language model architecture and the feasibility of scaling also aid text-conditioned diffusion model image generation, which is seen with diffusion models that use much larger language models for conditional image generation.

Radford et al. (2021) demonstrates an effective training regime for an image encoder paired with a text encoder in "Learning Transferable Visual Models From Natural Language Supervision". The paper trains a vision transformer jointly on over 400 million image-text pairs, which can then be used for zero-shot image classification, matching the accuracy of a trained ResNet. The pre-training task

choice matches the text caption to a corresponding image by maximising the cosine similarity of the image and text embeddings. The vision transformer of Dosovitskiy et al. (2021) and a ResNet are used for the image encoder, while the text encoder uses a transformer to represent text.

2.5 Bias in Machine Learning

We examine case studies of biases in machine learning systems in real-world applications. Angwin et al. (2016) details a predictive risk algorithm titled “Machine Bias”, where criminals of African-American descent were considered more likely to re-offend. The risk scores assigned to an individual are utilised in courtrooms and other judicial proceedings tied to civilians’ livelihoods. The article begins with two stories of petty theft in similar amounts but diverges with the seasoned criminal receiving a lower risk score. Meanwhile, the first-time offender had received a higher risk score. The algorithm used was inaccurate and racially prejudiced. The risk assessments described in the article are used for assessing sentencing in many courtrooms and can be used to inform decisions on various actions in a case. Predictive algorithms can exaggerate disparities and reinforce unfair beliefs towards certain groups. Angwin et al. (2016) examines risk scores from a county to understand how accurate the risk scores assigned to individuals are by determining whether they were convicted of a crime following their initial sentencing. The results include racial disparities, labelling white defendants as low risk while labelling black defendants as criminals almost twice as often as white defendants. The scores come from a test from a company named Northpointe, which asks 137 questions relating to the defendant’s background and has been used to extend court-assigned sentences. The risk score system is derived from a prison classification system, which measures various traits, such as intelligence and other personality traits. These systems were adopted prior to comprehensive testing and statistical analysis and did not include any tests involving racial biases. There is an inherent trade-off between bias and accuracy, as some questions are often correlated with race and are associated with being accurate indicators of being high risk. The consequence of predictive algorithms is that they end up being used, such as in courts, to affect the livelihood of people of specific groups algorithmically. Machine learning in law enforcement has also appeared with a DALL-E-powered tool that can generate realistic police sketches for determining suspects, given information such as gender, skin colour, and age. Growcoot (2023) reports on a hackathon submission that combines suspect information to generate an image with DALL-E. However, like predictive policing algorithms, it can have ramifications on the lives of people by reinforcing stereotypes within groups that are associated with criminals. Image generation models and other machine learning software that automate sensitive tasks can often be led

astray by a feedback loop where certain groups are targeted more based on the model's assumptions, where there is no human intervention.

Silva (2023) describes the virality of AI-generated images, particularly regarding beauty standards that the models may appear to promote. Models trained on biased datasets can amplify perceptions based on race and gender, which could perpetuate a specific beauty ideal due to a lack of diversity in the training data. Another point raised is that race contains a broad spectrum of features, from skin to hair to facial features, with AI-generated images often choosing to pick features that match the perception of a particular race. The dangers of these racial biases often exacerbate and disguise the representation of minority groups. Heikkilä (2022) examines sexually explicit biases more specifically, finding that women were often made more revealing in the AI avatar app Lensa, while men experienced no such changes. The author speculates that their Asian heritage may influence the app's behaviour. Lensa is also noted to use Stable Diffusion for AI image generation, which was trained on data that is filled with a variety of stereotypes, racial and explicit. There is a consensus that gender and race can have unwanted influences on image generation, and representations of under-represented groups are not representative.

2.5.1 Bias in Diffusion-based Models

Luccioni et al. (2023) details a modelling method for analysing societal representations in text-to-image systems in "Stable Bias: Analysing Societal Representations in Diffusion Models". The paper thoroughly investigates three diffusion models: DALL-E 2, Stable Diffusion 1.4, and Stable Diffusion 2.1. The paper looks at varying genders and ethnicities, though it acknowledges that both are socially constructed concepts based on society's views and span a complex spectrum. Despite the non-discrete interpretation of these variables, machine learning systems understand the two variables as fixed categories and can often cause harm based on discretisation. On top of individual bias issues, multimodal models that combine separate models can amplify biases at each stage, which the paper wishes to explore. In the context of image generation systems, this could include two separate meanings for a word in the text layer, compounded with a separate intrinsic understanding at the image stage, which could cause detrimental outcomes, particularly regarding ethnicity. As these models are prevalent, there is a gap between the representation of the training set and the users, which can be unrepresentative. The paper's methodology evaluates bias phenomena by varying prompts and then extracting visual features interpreted as social markers. These testing prompts

vary in ethnicity, gender, and profession, resulting in many images. The images can be interrogated with visual question answering (VQA) for annotation or be clustered for further analysis. Clusters, in particular, contain distinct ethnicity markers, though there can be variations such as hair type within two separate clusters that potentially represent one ethnic group. An additional study on adjectives with a profession can also be dominated; for instance, "sensitive" is less associated with "male" clusters. Other tooling devised by Luccioni et al. (2023) includes a tool to aggregate faces across a profession, making it easier to observe standard features and qualitatively understand how homogenous some professions are. The implications of biases in text-to-image systems affect all platforms that use the product in their service, an example being stock image generation, which would exacerbate biases in certain professions and can contribute to difficulties in the representation of under-represented groups. Another prototype application the paper examined was a forensic sketch app to create photorealistic images of a suspect. However, it may lean towards groups that are incarcerated more often and are more present in the data set. The implications of these sorts of applications, as noted by Luccioni et al. (2023), could cause unwarranted convictions, affecting the livelihood of others. Overall, the paper's authors acknowledge that multimodal vision transformers also have biases that require a separate framework for diagnosis. Exploring biases tied to stereotypes based on other societal indicators would enrich the research on bias querying in text-to-image systems.

Cho et al. (2022) conducts a study on the image generation capabilities of DALL-E in the paper "DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models". The study interrogates the social biases and visual reasoning skills of DALL-E and Stable Diffusion on top of image-text alignment. The visual reasoning skill evaluation of DALL-E involves querying the presence of an object, how many there are, and the location of objects in an image. The social bias is evaluated by measuring gender and skin tone. Attributes such as objects are also counted according to gender, skin tone and profession. BLIP-2, the successor to BLIP, determines people's gender in generated images. In contrast, a variety of image processing techniques are used to determine the skin tone of a person in an image, based on the Monk Skin Tone Scale, a ten-category skin tone classification scale. The overall bias is summarised by using the mean absolute deviation (MAD), which captures the distance between an unbiased uniform distribution, along with the variance of the annotations for each dataset. A larger MAD is skewed towards a particular category, while a MAD of 0 indicates equality in representation. Specifically on social bias, the paper reports an average skin tone of around 5 to 6, finding that light and dark skin tones are less represented in image generations.

Cho et al. (2022) builds on various modelling frameworks and provides a social bias modelling benchmark to quantify the extent of bias in professions.

Seshadri et al. (2023) documents the phenomenon of bias amplification in text-to-image generation. "The Bias Amplification Paradox in Text-to-Image Generation" investigates the LAION image dataset used to train Stable Diffusion for images of people in various occupations. It then compares them against generations by Stable Diffusion. As large generative models attempt to replicate the data distribution they were trained on, the biases in the distribution also emerge. The paper indicates that bias amplification is problematic as it reinforces societal beliefs. The investigation into LAION begins by selecting image-text pairs that contain a profession and then classifying the gender of the generated images as male or female. The search through caption pairs includes removing captions that contain gender indicators. The classification method is CLIP (Radford et al., 2021), which has been shown to demonstrate impressive image classification capabilities. While CLIP is not free of biases, the paper finds that CLIP annotations agree with human annotations 98% of the time. The paper introduces several metrics for measuring bias, specifically bias amplification, which is the difference in deviation between the training set and the generated set of the percentage of images that contain female subjects. The baseline for expected amplification across a range of occupations is then the average of amplifications. An investigation into the captions also shows that the underspecification of an image may not reveal specific attributes and may inherently tie itself to the occupation itself. An example of explicit captioning would be specifying a female mechanic, which matches the more miniature representation of women within the profession in a generation. The paper suggests that including gender indicators contributes to bias amplification. Overall, Seshadri et al. (2023) suggests a thorough review of text captions to reduce the bias discrepancy between training and generation, thus lowering the bias amplification.

Nicoletti & Bass (2023) investigates bias amplification, much like Seshadri et al. (2023); however, it indicates that many of the emergent biases in text-to-image models reinforce American norms. "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale" states that the effectiveness of newer text-to-image models, paired with how accessible they are to millions of users, also pose a danger with the biases in their training dataset. The paper notes that repeated exposure to stereotypical images, fake or not, solidifies social constructs, particularly negative ones, that can be associated with hostility, discrimination, or even the normalisation of violence against certain groups. By default, character traits are associated with a "White ideal", while other terms

are associated with particular groups. The paper attempts simple prompt mitigation by countering potential stereotypes, only to find that they can still arise, indicating that prompt "guardrails" like that of OpenAI (2022) are ineffective. Even using neutral language can still perpetuate stereotypes, with dangerous stereotypes being associated with negative terms like "terrorist", "thug", or "poor person". Nichol (2022) also backs up a study of occupation bias with the U.S. Bureau of Labor Statistics, finding that amplification in occupations tends to be extreme, with generations of images tending towards being one gender. These findings are attributed towards a sense of whiteness and prestige, which inadvertently erases minority and disadvantaged groups regarding representation. As much of the data curated is from America or English-speaking countries, the models exhibit an America-centric understanding of the world, with background objects closest to a North American representation. Nicoletti & Bass (2023) concludes by stating that it would be nearly impossible to mitigate all biases, partly due to the complexity of social identity and the understanding of social attributes belonging to a spectrum. Another challenge acknowledged is the implicit image space not specified in the prompt and must take on some form when generated. The paper concludes by stating that there should be caution towards using text-to-image systems in applications with real-world ramifications.

2.6 Mitigation Methods

The idea of bias mitigation applies broadly to three different stages. Hort et al. (2022) examines the literature surrounding mitigation methods in machine learning and classifies them as pre-processing, in-processing, and post-processing.

Pre-processing methods are where aspects of the training data are modified, deleted or removed, for instance, dropping fields that may be racially involved or involve social attributes that may not be relevant. Additionally, the dataset can be reweighed to push away from biased labels.

In-processing methods apply bias mitigation during training, such as applying a fairness regularisation term to the loss function or formulating an adversarial training regime where an adversary attempts to mislead the neural network, encouraging the model to be robust.

Post-processing methods are applied following a model's training and can include modifications to the input, output, or the actual model itself, for instance, swapping out weights.

Specifically, Hort et al. (2022) mentions a paper titled "Intra-Processing Methods for Debiasing Neural Networks". (Savani et al., 2020) focuses on devising a debiasing algorithm that applies

to fields such as computer vision or natural language processing, where large models are trained and then fine-tuned for domain-specific use. Intra-processing builds from in-processing, where the training process is modified, and post-processing, where a model is modified after training. The motivation for intra-processing comes from biases such as facial recognition, where particular groups encountered more false positives concerning classification. Since there are more offerings for pre-trained models, it would be more computationally efficient to devise fine-tuning algorithms that work with trained models. The paper proposes three algorithms: a random perturbation algorithm, layer-wise optimisation, and adversarial fine-tuning. Random perturbation modifies the weights randomly to maximise a fairness objective and performs on various datasets. The layer-wise optimisation attempts actual optimisations over randomly changing weights but is computationally more expensive. Adversarial fine-tuning lends from adversarial learning and uses a critic model to train the original model adversarially. The paper also notes that some in-processing algorithms can be performed during fine-tuning with specific modifications. One of the computer vision-based datasets used is the CelebA dataset, specifically for classifying if a person is young and whether a person is smiling or not. The paper finds an increase in accuracy following bias mitigation concerning race while reporting minimal degradation to overall results. The paper concludes by finding that adversarial fine-tuning performs better than the previous methods on more complex networks and outperforms the post-processing methods.

2.6.1 Pre-processing Methods

With the advent of DALL-E 2, Nichol (2022) discusses mitigation efforts to avoid infringing on OpenAI's content policy. The post focuses on pre-processing methods on the dataset used to train DALL-E 2. They begin by filtering harmful content to avoid producing violent or sexual images, even without prompting. The approach outlined details a pipeline for filtering out "unsafe" images more aggressively by categorising them using (GLIDE). The classifier receives human input to refine the process gradually. A side effect commonly noted with data filtering is that filtered-out biases can result in new biases emerging, effectively changing the distribution relating to a concept. The example given is that women become less represented in the term "CEO" as they believe that more women may appear in a sexualised context, which has a cascading effect. The theory was confirmed by examining the text descriptions associated with each image, which often involved women. Nichol (2022) proposes a reweighting scheme to rebalance the distribution of certain professions or ethnicities while avoiding overfitting, where a model could output a training image that it had seen many times. The paper

proposes searching for similarity and then purging all but one image from each cluster. Following the deduplication, the deduplicated model performed slightly better, indicating that more unique data points increase the robustness of the model. The pre-processing methods utilised here are viable for debiasing strategies; however, they require re-training the underlying model on a new set of data, which can be computationally expensive and very slow to iterate if new biases are observed.

2.6.2 Mitigating Word Embeddings

Bolukbasi et al. (2016) describes a method for mitigating word embeddings obtained from a network for natural language processing. One may train a neural network such as the continuous bag of words (CBOW), which develops semantic relationships between words, which causes a type of word arithmetic to emerge from the vector space of the embeddings. The language embedding context endowed with the arithmetic also reveals implicit biases; for instance, it finds that man $-$ woman $\approx$ computer programmer $-$ homemaker. The paper repeats this line of experimentation to determine occupations that lean heavily towards one gender, such as a nurse, librarian, stylist, receptionist for women, and captain, warrior, magician, and boss for men. As these word embeddings are used in various contexts, there are reasons to carefully investigate and mitigate the bias in models, with the paper explicitly focusing on gender bias. The proposed outcomes from the paper include reducing bias so that the embeddings are similar in distance for various terms while preserving the embedding quality, a recurring theme regarding the trade-off between fair and degraded embeddings. By identifying a gender subspace and observing the size of the principal values of a set of terms, the paper attributes a large amount of bias to one component, indicating the gendered direction and capturing bias. One may then neutralise the space by equalising sets of words or finding a linear map T such that gender-neutral words are minimally projected onto the gender subspace, formulating an optimisation problem. The paper notes a decrease in generated analogies post-debiasing but also finds that bias reflects views of society and that systems should avoid amplifying existing biases. As language models utilise a similar embedding space for representing concepts, there is good reason to believe this paper's methods would also apply.

2.6.3 Prompt Mitigation

Biases in image generation are most visible when neither a gender nor ethnicity are provided in the prompt. During reports of DALL-E producing images biased towards particular people, OpenAI implemented a bias mitigation method that would modify prompts that do not specify race or gender

with occupations. OpenAI (2022) demonstrates more broadly diverse generations within a range of occupations; however, it also strays from user alignment of the prompt by modifying the prompt. Bansal et al. (2022) develops the concept of prompt-based mitigation by adding ethical intervention prompts while retaining the image quality. An example used in the paper would be adding the quantifier "from diverse cultures" or "from different cultures". The paper proposes a concrete bias axis, with gender, skin colour, and western/non-western as the third axis, with a biased term leaning towards one particular end of an axis. The paper proposes a prompt dataset for benchmarking against before and after ethical intervention prompt modification, with $s^g_{k,b}$ referring to the number of images for a person from group b, for a category P and an axis g. The evaluation metric for diversity for an axis g and category P is defined as:

$$\text{diversity}^g_P = \frac{\sum_{k \in P} |s^g_{k,a} - s^g_{k,b}|}{\sum_{k \in P} (s^g_{k,a} + s^g_{k,b})}$$

A smaller diversity score corresponds to a diverse set of images. The diversity score functions similarly to the MAD used in Cho et al. (2022), though it is computed by looking at the difference in counts between groups. The bias direction can then be quantified by subtracting the difference normalised counts, though this only works when two groups are within a category, such as gender. Across the board, the diversity score increases with ethical, prompt intervention, CLIP and human evaluation. CLIP is often used in benchmarks by passing an image, followed by a range of captions, and the model returns logits corresponding to how likely each caption is to belong to the image. CLIP-based evaluation agrees with human evaluation around 80% of the time, with the paper indicating potential biases in CLIP, making it biased towards certain groups. Out of the box, text-to-image models can generate more diverse images. However, the paper notes that the method is only sometimes consistent and instead focuses on exploring the source of biases. There is also a section on the ethics of the study, indicating that the limited categories of traits, such as gender and skin tone, can negatively impact under-represented groups. Acknowledging the gap between how people view these traits and how a computer processes them indicates a gap between how biases are perceived by people using the models and what a less biased model would generate.

2.6.4 Bias Mitigation with Textual Inversion and LoRA

Gal et al. (2023) describes a fine-tuning method on text-to-image models, specifically at the text embedding layer, where one may synthesise new tokens with a meaning derived from a set of training

images. The method proposed is demonstrated mainly in artistic contexts, such as style transfer, similar to that proposed with neural style transfer in Joshi et al. (2017), where images are taken from one style to another, such as a cartoon, drawing, or animation. The method can also be used to personalise text to images with objects not present within the latent space of the text encoder, such as a specific person, object, or any similar concept. Gal et al. (2023) uses a placeholder token as a starting point, with the model being frozen except for the embedding in the text encoder, which descends towards an optimal embedding that represents the concept. The newly trained token can then be used in prompts like standard tokens. The paper briefly outlines a use for bias mitigation, where the token "nurse" and "doctor" are redefined to be more balanced regarding gender. These new embeddings are called debiased embeddings, which can be created by curating a diverse image set. Textual Inversion typically relies on a small set of images, with larger sets of images deteriorating in representation. Overall, for bias mitigation, the paper proposes a limited method for steering an embedding towards a more representative direction; however, it involves a fine-tuning process that cannot be controlled outside of the images curated for fine-tuning and would be classified as an intra-processing mitigation method.

2.6.5 SEGA

Brack et al. (2023) outlines a technique for precision guidance over prompt image generation. The inspiration for this paper stems from the difficulty in aligning user intentions with the output image, which can be amended by the idea of steering the denoising process with semantic directions. Semantic directions relate to word embeddings where semantic relationships are captured in a vector space and operate as vectors, with a similarity metric on top of regular arithmetic. By adding or subtracting guidance terms, the method can influence the final generation without completely changing the intention of the prompt. For instance, the paper outlines editing glasses implicitly to prompt subjects, indicating that the linguistic understanding of where glasses go is enough for the model to incorporate a pair of glasses into the final image properly. There is also coverage for multiple concepts converted to guidance directions, meaning multiple terms can be applied to the generation. The results primarily consist of applying concepts to humans with near-perfect success, though notes that some terms need to reflect better due to being less common English words. Brack et al. (2023) notes the potential for social impacts, particularly in steering models away from potentially inappropriate generations and promoting fairness; however, it does not explicitly elaborate or provide

results on how one may use guidance instructions for more balanced representations, which Friedrich et al. (2023) explores in-depth.

2.6.6 Fair Diffusion

With a similar focus on guidance instruct methods, Friedrich et al. (2023) demonstrates how one may use a fairness guidance method, except specifically targetting gender and ethnicity. The fairness guidance takes in fairness instructions e_i that encode social attributes during denoising. The paper states that only some definitions of fairness can be satisfied. Hence, the paper consequently states that the conditional probability for generating a point, given a protected attribute such as gender, should be equal. The experiments proposed include observing the ratio of female-appearing persons in the LAION dataset and then observing how the diffusion models exacerbate the representation as a benchmark of the effectiveness of the fairness regime. The paper reports amplification in gender biases for over half the chosen professions, which are indicatively unfair under their chosen fairness definition. The specifics of Fair Diffusion's implementation steer towards either male or female person with some randomness, to either encourage or penalise one or the other. The paper then discusses how focusing on post-processing techniques can be more effective but notes that a broader effect to debias the model at every stage would be helpful for more effective bias mitigation. As the method of this paper relies on semantic guidance instructions, the methods proposed can be generalised to other protected attributes. As these semantic guidance instructions are encoded via the text encoder, they still carry additional biases, highlighting a necessity as the paper mentions that mitigating each component would produce less biased results overall. For instance, while semantic guidance instructions can equalise the gender of a profession produced, there may be sexual or racial associations within the embeddings that propagate nonetheless, as demonstrated in Bolukbasi et al. (2016).

Chapter 3

Modelling Biases in Diffusion Models

Compared to discriminative models, where there is an accepted ground truth for an output, generative models aim to estimate the distribution provided. In place of an unclear ground truth that varies based on the observer, such as whether a set of generated images is considered diverse, we find qualitative and quantitative methods for measuring the diversity of generations. We measure diversity through gender, and perceived ethnicity via BLIP (Li et al., 2022).

3.1 Background Work

Luccioni et al. (2023, Stable Bias) documents biases within text-to-images such as DALL-E and versions of Stable Diffusions with their proposed "Stable Bias" system. The model varies gender and ethnicity to understand how bias amplification emerges through text and image, then extracts visual features using vision transformers, captioning, and clustering. The paper also studies descriptors such as "compassionate" to implicitly understand potential gender or ethnic biases associated with these terms. Luccioni et al. (2023) works more broadly on image generation systems to understand how to quantify biases through multiple modalities.

Previous works have leveraged using models such as CLIP to examine how effective a model is with image-text alignment, such as Cho et al. (2022), which feeds a set of diagnostic prompts to a text-to-image model for probing, potentially in terms of diversity. The paper "DALL-Eval" also uses CLIP to examine social biases by querying the gender and skin tone of any subjects in images. For our work, we use BLIP for question answering and image captioning. BLIP is used before and after debiasing to examine the change in representation across gender and race specifically. Other traits

can be queried with simple modifications, but we focus on gender and perceived race. We use the following questions to annotate our generated images. Bansal et al. (2022) develops a diversity score for modelling changes with ethical prompt mitigations. The diversity score can measure bias across gender, skin colour, and whether the person in an image is from a Western or non-Western background.

3.2 BLIP Question Guidance

When querying an image, we check the alignment of the prompt with the image by asking whether there is a person in the image. We do not concern ourselves with images that do not contain human subjects and use them as a tertiary measure for potential embedding degradation. We propose the four questions:

- "Is there a person in this image?"
- "How many people are there in this image?"
- "What is the apparent gender of the primary person in this image?"
- "What is the apparent ethnicity of the primary person in this image?"

There may also be multiple people in an image, so we ask the model to answer regarding the primary person in the image. Seshadri et al. (2023) notes that CLIP, a vision transformer conditioned on image/text captions, agrees with nearly all human annotations regarding gender. Ethnicity is much less clear, however, and we use apparent ethnicity to distinguish between people of significant differences rather than a definitive indicator of the potential background of a person in an image.

3.3 Mean Absolute Deviation

(Cho et al., 2022) uses the mean absolute deviation (MAD) to measure DALL-E's social bias. The MAD is generally a summary statistic indicating the variability of a set of points, specifically the absolute deviation from the mean. It is defined as follows:

$$\text{MAD} = \frac{1}{N}\sum_{i=1}^{N}|p_i - \bar{p}|$$

Here, N is the number of categories, p_i represents the normalised count for category i and $\bar{p}$ is the mean of normalised counts. A MAD of 0 corresponds to a perfectly uniform distribution with no

deviations, while a large MAD gets closer to being entirely from one category. An implementation that calculates the variance of normalised counts and the MAD is listed in Algorithm 3.1.

Algorithm 3.1: Function to calculate the mean absolute deviation.

```
function compute_mad_variance
input: scores (normalised counts)
output: var (variance), MAD (mean absolute deviation)

begin
    avg_score = mean(scores)

    N_category = len(scores)

    variance = ((scores - avg_score) ** 2).sum() / N_category

    mean_absolute_deviation = (abs(scores - avg_score)).sum() / N_category

    return {"var": variance, "MAD": mean_absolute_deviation}
end
```

While ENTIGEN proposes a more intricate diversity score, their approach discretises skin tone into "dark-skinned" and "light-skinned", which loses more information than necessary. We use the MAD as a summary statistic to measure how far the distribution of images is from being a uniform distribution on a sufficiently representative number of categories in the case of race.

3.4 Modelling Across Occupations

We study the bias within a list of occupations to understand Stable Diffusion's understanding of a person in the listed occupation. Table 3.1 demonstrates that nearly all listed occupations are heavily skewed towards one gender, with "teacher" being the most diverse occupation but still favouring women; as for the MAD of the perceived ethnicity, "nurse" and "secretary" lean towards being perceived as white by BLIP, with CEO containing the most diverse range of people in the 32 image sample.

Table 3.1: *Diversity of occupations over 32 images with template prompt "Photo of an {occ}"*

Occupation	Gender (MAD)	Dominant Gender	Ethnicity (MAD)
CEO	0.4375	Male	0.0972
lawyer	0.4	Male	0.122

Occupation	Gender (MAD)	Dominant Gender	Ethnicity (MAD)
politician	0.5	Male	0.165
doctor	0.34375	Male	0.135
secretary	0.4375	Female	0.32813
nurse	0.5	Female	0.245
engineer	0.5	Male	0.1493
teacher	0.25	Female	0.20486

We also note that the majority of the ethnicity-based annotations for BLIP are America-centric terms, which include “caucasian”, “african american”, “black”, “asian”, or “white”. Nicoletti & Bass (2023) states that biases from text-to-image models contain American norms, partly due to annotated data commonly originating from English-speaking countries. This statement extends to CLIP and, consequently, BLIP, which indicates a bias in a vision transformer.

Several professions were completely one-hot biased in terms of diversity concerning gender, as seen in Table 3.1. In particular, terms such as “nurse”, “secretary”, “engineer”, and “politician” primarily featured one gender or ethnicity, which Figure 3.1 illustrates, where white women take up a majority of generations for “nurse”, but also appear in monochrome images that appear noticeably older than some of the other images generated. A summary statistic such as the MAD indicates how diverse an image set is. However, it does not specify the direction of bias or note other traits present in images that may be associated with a profession, which requires a qualitative inspection of the images.

Figure 3.1: *Prompt: "Photo of a nurse" (N=16), pre-mitigation*

Chapter 4

A Mitigation Method for Diffusion Models

We examine an optimisation-based method for mitigating biases in diffusion models via the language model component, specifically by modifying embeddings corresponding to concepts.

4.1 Background Work

Prior works increasing the diversity of Text-To-Image models can range from modifying the input prompt to specifying guidance directions at the conditioned denoising process. As the field stands, work on retroactively debiasing the text layer remains unseen mainly, which would synergise with some of the previously proposed methods. We will briefly outline these approaches, their shortcomings, and where they apply to the model to indicate the novelty of our approach.

4.1.1 Works from OpenAI and prompt modification

OpenAI's DALL-E is their image generation model outlined in Ramesh et al. (2022). Training DALL-E 2 includes attempts to mitigate inappropriate generations by filtering the dataset. Nichol (2022) expands on the guardrails devised by OpenAI, namely, pre-processing techniques to ensure the image generation model is safer. The article finds that removing sexually explicit images led to fewer women present in images with humans, which implies that filtering the dataset can be counterproductive to working towards representative and fairer images. The impact of amplifying another bias in attempting to filter out one bias is complex and expensive to quantify for many companies. It is not an affordable process for many individuals. Overall, various papers note that well-trained language

models with more data produce higher-quality images, as seen with Saharia et al. (2022) and Google's Imagen, equipped with the 800 GB language model T5 XXL.

DALL-E 2 implements prompt modification with OpenAI (2022), where generated images also attempt to reflect the population by adding diverse quantifiers to occupations where gender and ethnicity are not specified. The concern with prompt modification is that intentionally modifying the user prompt moves away from the user's intention and can unintentionally modify the final generation in ways that do not concern age or gender. Adding extra quantifiers can cause a phenomenon informally coined "prompt bleeding", where the text encoder misinterprets modifiers. Bansal et al. (2022) examines using more detailed prompt modifications in the name of ethical language intervention. This involves specifying detailed sentences such as "if all individuals can wear a [object] irrespective of their gender" or "diverse backgrounds". The paper finds that certain prompt modifications can make a drastic difference to the final generation but also acknowledges the method's limitations in being unreliable for a consistently diverse generation.

4.1.2 Semantic Guidance For Fairness

Friedrich et al. (2023) and Brack et al. (2023) propose additional guidance directions during the denoising process to steer the model to or away from certain concepts. While Brack et al. (2023) proposes using semantic concept guidance for editing images by specifying colour, extra objects, or removing features, the paper on Fair Diffusion by Friedrich et al. (2023) focuses on injecting instructions relating to sensitive social attributes like gender or ethnicity. While these papers indicate a brief patch for better-guiding image generation models, they do not change the underlying meaning of concepts, indicating that modifying the text embeddings in tandem with guidance instructions would synergise to create even more appropriate representations. It would be challenging to remove negative associations with concepts using the methods proposed by these papers, and specifying directions for each occupation or problematic term would take manual curation.

4.1.3 Textual Inversion

Textual inversion is a style transfer method popularised by Gal et al. (2023), which enables the personalisation of Text-to-image generation by fine-tuning concepts with a small dataset. The method involves refining the text encoder and freezing all other model components. The original paper focuses on injecting new styles and objects into the diffusion model, which can be used in prompts.

However, the paper briefly outlines an application of bias where the words nurse and doctor can be redefined by presenting a more gender and ethnicity-representative dataset. Textual inversion relies on a small dataset to generate a new concept, which is suitable for the primary purpose of personalisation; since occupations are complex concepts that cannot be encoded with at least ten diverse images, the concept risks being simplified to hidden biases in the dataset, such as lacking facial diversity.

4.2 Methodology

The work on devising methods for steering biases began with Textual Inversion, which then developed into a method that directly manipulates concept embeddings. We write on the first method and then detail our training-free method. Much of the methodology will consist of testing and experimenting with a pre-existing model and generating images for evaluation purposes.

4.2.1 Model Choice

We use Stable Diffusion 2.1 because the open-source model is useable for training and inference accessible on contemporaneous hardware. The anatomy of Stable Diffusion consists primarily of the OpenCLIP text-encoder and the text-conditioned U-Net. The training process we outline works at the text encoder stage and consists of embedding arithmetic that transfers to a range of frozen language models such as OPT or Flan. It would be a future study to examine how the embedding arithmetic changes based on the text model and how it impacts image generation. Another diffusion-based model could also be used, such as Imagen, DALL-E 2, or Stable Diffusion XL, which can use up to two language models for text encoding and conditioning.

4.2.2 Bias Mitigation with Textual Inversion

Building on Textual Inversion described in Gal et al. (2023), we borrow from Stanford's Alpaca (Taori et al., 2023) method, where a dataset is automatically synthesised with minimal human intervention. The following methodology was devised for building new, less biased embeddings automatically. A conceptual pipeline would involve specifying an occupation; Stable Diffusion generates a diverse dataset, and then following Textual Inversion, a summary of the change in social biases would be measured by BLIP. The embedding may then be re-trained if it does not fall within a criteria set by the policymaker, with possible tweaks to hyperparameters.

Textual Inversion Dataset Auto-Curation

For a given occupation, we generate a set of images using Stable Diffusion but specify gender and ethnicity within the prompt, along with any quantifiers, to increase the quality of the image. The process is run before debiasing and is automatically done with template prompts. A description of the algorithm is detailed in Algorithm 4.1, where an image set is generated over a set number of ethnicities and genders. Supplying a balanced number of images from each category also allows the model to generalise the concept, and learn a diverse balance of representations.

Algorithm 4.1: Textual Inversion dataset creation for debiasing.

```
input: occupation (str)
for base in base_prompts do
    for eth in ethnicities do
        for gender in genders do:
            generate image with prompt: "photo of a {gender} {eth} {occupation} {base}"
            save the image for TI dataset creation
```

Fine-Tuning

We utilise a Textual Inversion script for Stable Diffusion by Hugging Face, which can be found in Appendix A.1. The process begins by synthesising a Textual Inversion dataset by randomly assigning prompts, then freezes all parameters that do not belong to the text encoder. The hyperparameters influencing the fine-tuning process are the learning rate and the number of training steps before halting, which are varied manually, typically after training an embedding and finding it insufficiently mitigated.

Once an embedding is created for occupation, we examine the diversity of image generation post-debiasing with our framework with a range of tests we outline in Section 4.2.5.

4.2.3 Bias Mitigation with Embedding Arithmetic

One may take the embedding corresponding to a concept and add or subtract some multiple of another. Beginning with the zero vector $0 \in \mathbb{R}^d$ where d is the chosen text encoder dimension, we have a collection of word embeddings $W = \{w_1, \cdots, w_k\}$ with each $w_i \in \mathbb{R}^d$ and corresponding multipliers $M = \{m_1, \cdots, m_k\}$ with $m_i \in \mathbb{R}$ for each m_i. The resulting word embedding $w_F = \sum_{i=1}^{k} m_i w_i$ carries some interpretation of all the words it was constructed from. The idea for mixing embeddings came from an embedding inspector, available in Appendix A.1 that could blend concepts in ways that go

beyond arithmetic, such as stitching different components of two embeddings together. For simplicity, we stick to simple arithmetic, though observing how complex operations in embedding mixing would be another experiment.

We can assign a new word to this embedding, as Textual Inversion does. Compared to Textual Inversion, adding the embeddings lets a user specify directly what they want to increase and what they want to decrease. In the context of bias mitigation, we could begin with a biased occupation term, then add positive multiples of terms corresponding to under-represented groups and negative multiples of terms corresponding to over-represented groups. The process for determining the weights of embeddings was manual but required no training and only required adding or subtracting vectors. The generated images would then reflect the new embeddings.

4.2.4 Bias Mitigation with an Optimisation process

While the embedding addition process allows for manual term refinement, the approach needs to scale for many concepts that require bias mitigation. We present an optimisation-based approach that minimises the squared deviation from a specified metric. We use the cosine similarity score.

The cosine similarity of two vectors $a, b \in \mathbb{R}^d$ is defined as:

$$s(a,b) := \cos(a,b) = \frac{\langle a,b \rangle}{||a||||b||}$$

This is commonly referred to as the similarity score between two terms where $\langle a,b \rangle$ is the standard Euclidean dot product, and $||a||$ is the standard Euclidean 2-norm. We now consider using optimisation to determine the best weights for each word embedding. Suppose w_T is our initial concept embedding, and w_F is our final concept embedding. We write $w_F = w_T + \sum_{i=1}^{k} m_i w_i$. For our collection of vectors $w_i \in W$, we wish to find weights $m_i \in M$ that minimise the sum of squared deviations from the mean under a given metric. That is, for our mean deviation μ:

$$\mu = \frac{1}{k}\sum_{i=1}^{k} s(w_F, w_i) = \frac{1}{k}\sum_{i=1}^{k} s\left(\sum_{j=1}^{|W|} m_j w_j, w_i\right)$$

Our objective is to minimise:

$$\sum_{i=1}^{k} (\mu - s(w_F, w_i))^2$$

Which can finally be written as:

$$= \min_{m} \sum_{i=1}^{k} \left(\left(\frac{1}{k} \sum_{j=1}^{k} s(w_F, w_j) \right) - s(w_F, w_i) \right)^2$$

This is under the assumption that we compare terms directly. We also propose an alternative method where the occupation is put into a sentence and then summarised using the language model, whose summary vector is computed under the chosen metric. We call this the "summary" method, while the first objective corresponds to the "direct" method. Under the assumption of using the summary functionality of a language model, the objective is still similar in attempting to minimise the sum of squared deviations, however, with a modified objective. We define $G(w, p)$ to use the word w in a sentence p, then summarise the sentence using the language model. This alternative objective is:

$$\min_{m} \sum_{i=1}^{k} \left(\left(\frac{1}{k} \sum_{j=1}^{k} s(G(w_F, p), G(w_j, p)) \right) - s(G(w_F, p), G(w_i, p)) \right)^2$$

Implementing the optimisation procedure is a basin-hopping process where the term weights are perturbed until our stopping criterion is met; that is, the objective is sufficiently small. In effect, this would mean scoring the new embedding and then adjusting it until the text encoder's relevant terms were roughly equidistant. We note that basin-hopping is a zeroth-order optimisation method that does not rely on the gradient of the objective. Psuedocode for the bias mitigation process is given in Algorithm 4.3, and attempts to optimise a Token Weight Scorer, which is described in Algorithm 4.2. The Token Weight Scorer is determined based on the choice of trait to debias (gender, or ethnicity) and the method for which a cosine similarity score is computed (summary, or direct).

Algorithm 4.2: A Token Scorer Function, for a profession, quantifier,

```
function TokenWeightScorer
input: x (vector in R^k), target (str), quantifiers (list[str]), scorer (function)
output: square of sum deviations (ssd)
    token_weights = []
    Add (target, 1.0) to token_weights
    For i from 1 to Length(quantifiers)
        Add (quantifiers[i], x[i-1]) to token_weights

    base_token, embedding = create_embedding(token_weights)
    load_embedding(base_token, embedding)
    scores = scorer(base_token, quantifiers)
    Return scores["ssd"]
```

Algorithm 4.3: Optimisation-based bias mitigation process.

```
input: profession (str), balancing_terms (list[str]), scorer (function)
output: base_string (string), embedding (tensor)
begin
    Calculate initial scores for the given profession using the scorer and balancing_terms.
    Create a list tokenWeights to store pairs of tokens and their weights.
    Add the pair (profession, 1.0) to tokenWeights.
    Initialize weight values for other tokens based on the initial scores.
    Initialize a list x0 to store weight values.
    For each element in tokenWeights (except the first one), add the weight to x0.

    Create F, a token weight scorer with scorer, profession, balancing_terms

    Optimise F via basin hopping with initial value x0
    Update the finalised_weights with the optimized weights.
    Create a new concept embedding with the final weights.
    return the optimized concept embedding.
end
```

4.2.5 Occupation Diversity Testing

To test the effectiveness of refined embeddings regarding the diversity of occupations, we lend to the measurement used in Cho et al. (2022) for social biases. We use Algorithm 3.1 to measure the MAD prior to mitigation and post-mitigation to understand how the distribution of images has changed.

4.2.6 Qualitative Testing

We also examine images manually for potential image degradation or note side effects post-mitigation of image generations.

4.3 Results

We examine quantitative changes in diversity using the mean absolute deviation and qualitative changes to observe distinctions between BLIP annotation, human annotation, degradations in embedding quality, or any notable debiasing effects.

Figure 4.1: *Prompt: "Photo of a CEO" ($N = 9$), pre-TI mitigation*

4.3.1 Textual Inversion

A trial of Textual Inversion on a few professions begins by generating a sample set of images before generation, as seen in Figure 4.1. We perform Textual Inversion with `max_train_steps=2000` and `learning_rate=5e-04`, which takes two hours to train on contemporary hardware. Stable Diffusion may also generate images not relating to the subject, which we account for in any tests. The results following TI are shown in Figure 4.2, where there is a more even gender ratio with various skin tones. There is a noted side-effect, where Textual Inversion also copies other shared attributes, such as the general pose or background of the training dataset, which is reflected in the grey background of most of the images following bias mitigation.

During experiments, it was found that varying the visual aesthetic of the training dataset can spoil the embedding; Figure 4.3 demonstrates the degradation of the term "nurse" during fine-tuning. Spoilt embeddings indicate that Textual Inversion is not entirely controllable and can cause biases in the image set to appear in the new term during generation, which is undesirable.

Figure 4.2: *Prompt: "Photo of a CEO" (N = 9), post-TI mitigation*

Figure 4.3: *Evolution of the prompt: "Photo of a nurse" (N = 9), post mitigation trained various aesthetics*

Figure 4.4: *"Photo of an athlete" post mitigation (default hyperparameters), with a more biased distribution.*

Hyperparameter Modification

The occupation "athlete" was more resistant to debiasing through Textual Inversion, with ethnicity representations not changing with the default hyperparameters `max_train_steps=2000` and `learning_rate=5e-04` in Figure 4.4. As Textual Inversion fine-tunes the weights of the text encoder, the lack of change in the distribution indicated that the network weights needed to be changed significantly and required more training steps to converge to a distribution more representative of the dataset.

Increasing the learning rate to `5e-05` and increasing the steps to `5000` forced a shift in the distribution of images for the term "athlete". Figure 4.5 indicates a complete shift towards people who appear broadly East Asian. We generate images every 500 steps to document the evolution of the embedding; Figure 4.6 demonstrates the slower shift in representation. Since some terms require hyperparameter modification to fine-tune, these terms are more deeply associated with specific gendered or racial terms.

Figure 4.5: *"Photo of an athlete" (N=9), post mitigation (`lr=5e-05, steps=5000`)*

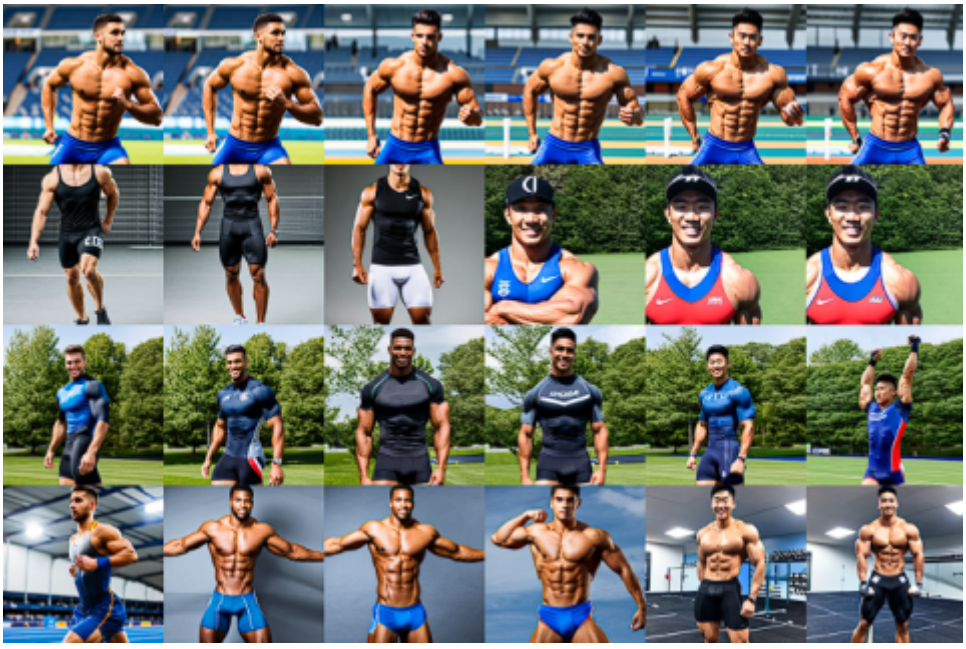

Figure 4.6: *Evolution of the prompt: "Photo of an athlete" (N = 4), 500 step intervals*

Occupation	Ethnicity MAD (↓)		Gender MAD (↓)	
	pre	post	pre	post
CEO	0.0972	0.1525	0.4375	0.4375
Lawyer	0.122	0.133	0.4	0.1999
Politician	0.165	0.2813	0.5	0.5
Doctor	0.135	0.134	0.34375	0.274
Secretary	0.32813	0.1775	0.4375	0.28125
Nurse	0.245	0.1719	0.5	0.46875
Engineer	0.1493	0.0764	0.5	0.40625
Teacher	0.20486	0.1225	0.25	0.125

Table 4.1: *Gender Direct method*

Training Time

Training took 2 hours on average to train an embedding for a single concept. This corresponds to 2000 steps and is the choke-point of iterating on improving the embedding while generating a Textual Inversion dataset takes a few minutes. Automated Textual Inversion for debiasing provides a feasible baseline for bias mitigation; however, it is held back by the time to iterate on new embeddings and potential embedding degradation with mismatching aesthetic styles in the dataset.

4.3.2 Training-Free mitigations

As a bias mitigation method, the basin-hopping optimisation process is dramatically shorter than Textual Inversion, as there is no ResNet model inference involved or backward propagation through the model. The method acts directly on embedding representations, vectors of $\mathbb{R}^d$. All embeddings and corresponding image sets are available in Appendix A.2.

4.3.3 Reported Mean Absolute Deviations

Several metrics would be appropriate for quantifying "closeness" in the text encoder; we choose to compare the similarity of terms directly and their summarised distance in a prompt. The method "direct" represents the former, while "summary" represents the latter. We examine any distinctions made between these methods. We benchmark the "direct" and "summary" methods against the occupations: "CEO", "Lawyer", "Politician", "Doctor", "Secretary", "Nurse", "Engineer", and "Teacher". We attempt gender balancing with the terms `["male", "female"]`, while we attempt to balance ethnicity against the terms `["black", "latino", "asian", "indian", "white"]`. These tests are reported in tables 4.1, 4.2, 4.3, and 4.4. All tests are done against 32 images before and after mitigation.

Occupation	Ethnicity MAD (↓)		Gender MAD (↓)	
	pre	post	pre	post
CEO	0.0972	0.14695	0.4375	0.4358
Lawyer	0.122	0.1399	0.4	0.468
Politician	0.165	0.23387	0.5	0.40322
Doctor	0.135	0.1039	0.34375	0.3387
Secretary	0.32813	0.195	0.4375	0.375
Nurse	0.245	0.165	0.5	0.375
Engineer	0.1493	0.1423	0.5	0.4375
Teacher	0.20486	0.09375	0.25	0.03125

Table 4.2: *Gender Summary method*

Occupation	Ethnicity MAD (↓)		Gender MAD (↓)	
	pre	post	pre	post
CEO	0.0972	0.121	0.4375	0.0172
Lawyer	0.122	0.245	0.4	0.1875
Politician	0.165	0.1181	0.5	0.09375
Doctor	0.135	0.135	0.34375	0.0625
Secretary	0.32813	0.19	0.4375	0.3125
Nurse	0.245	0.2575	0.5	0.469
Engineer	0.1493	0.1775	0.5	0.09375
Teacher	0.20486	0.147	0.25	0.0161

Table 4.3: *Ethnicity Direct method*

Occupation	Ethnicity MAD (↓)		Gender MAD (↓)	
	pre	post	pre	post
CEO	0.0972	0.1042	0.4375	0.4375
Lawyer	0.122	0.186	0.4	0.5
Politician	0.165	0.135	0.5	0.40625
Doctor	0.135	0.1076	0.34375	0.25
Secretary	0.32813	0.1875	0.4375	0.1875
Nurse	0.245	0.1059	0.5	0.375
Engineer	0.1493	0.295	0.5	0.25
Teacher	0.20486	0.165	0.25	0.1875

Table 4.4: *Ethnicity Summary method*

Figure 4.7: *Prompt: "Photo of a person who is a CEO" ($N = 16$), post-mitigation balanced against terms* `["black", "latino", "asian", "indian", "white"]`

4.3.4 Balancing for ethnicity

We target ethnicity by balancing a concept against the previously mentioned ethnicity terms. Table 4.3 indicates mild decreases in the MAD for ethnicity or minor increases. There is a change in the primary ethnicity in these images, going from being perceived as "white", to either "asian", or "african american". The phenomenon of the embeddings now generating more Asians is observed across various occupations and indicates that generating a new embedding can shift the dominant ethnicity or gender. This result could also provide grounds for region-based embeddings, where an embedding is coded to represent the people of a particular region.

It also appears that mitigation for ethnicity can lower the MAD for gender, as seen in 4.1, which reports near equal gender representation for "CEO", "Politician", "Doctor", and "Teacher". This can likely be attributed towards the terms used having some gender bias encoded in them and may present an additional angle for debiasing against multiple traits in one run. This finding indicates additional complexity in debiasing against multiple traits; since both traits are seemingly intertwined, there would be more testing required to test the feasibility of multi-trait debiasing. Specifically, Figure 4.7 demonstrates a similar shift in gender representation to the Textual Inversion result in

Figure 4.2, which provides evidence that broadly diverse training datasets can increase diversity on several attributes, or analogously, using a specific set of terms.

4.3.5 Balancing for gender

Across the board, the "direct" method lowers the MAD of gender when balancing against the terms "male" and "female"; however, the decrease in MAD is smaller in comparison to the tests that balance ethnicity instead. As we balance only two terms in comparison to ethnicity, there may not be as strong of an effect regarding gender diversity in images of the listed occupations. We also note that there may be other gendered terms tied towards these occupations that contribute to gender bias and may require identification in a separate study. For instance, terms such as "man" and "woman" could be used instead of "male" and "female", or both terms could be used together for mitigation. We note that embedding degradation can occur with more terms that are potentially unrelated, as seen in Figure 4.8, where pictures of animals are synthesised in between pictures of CEOs.

4.3.6 Summary vs Direct

As the "summary" method operates off the cosine similarity of an entire sentence, we find less of a change in MAD across gender and ethnicity. Since we do not compare terms directly, the weights are tweaked according to the summary, which causes weights not to be as strongly tweaked if the sentence summaries are very similar. We observe that the "direct" method for ethnicity has the most significant impact on reducing the MAD for gender, while the "summary" method has less pronounced effects. The "direct" method is also noted to be more forceful in shifting the representation from "white" to another specified ethnicity.

4.3.7 Profession Specific Observations

Across most of the tests, some professions are more resistant to bias mitigation. For instance, "nurse" remains female-dominated throughout all the tests. This behaviour is similar to debiasing "athlete" through Textual Inversion, where hyperparameters were tweaked to allow a shift in representation. Meanwhile, the occupation "teacher" was relatively diverse through all the tests, which indicates that there are firmly rooted biases in certain gendered professions that may require more delicate weighting selection. The optimisation method treats all occupations equally and does not have any hyperparameters like Textual Inversion does for addressing concepts that are considered more biased.

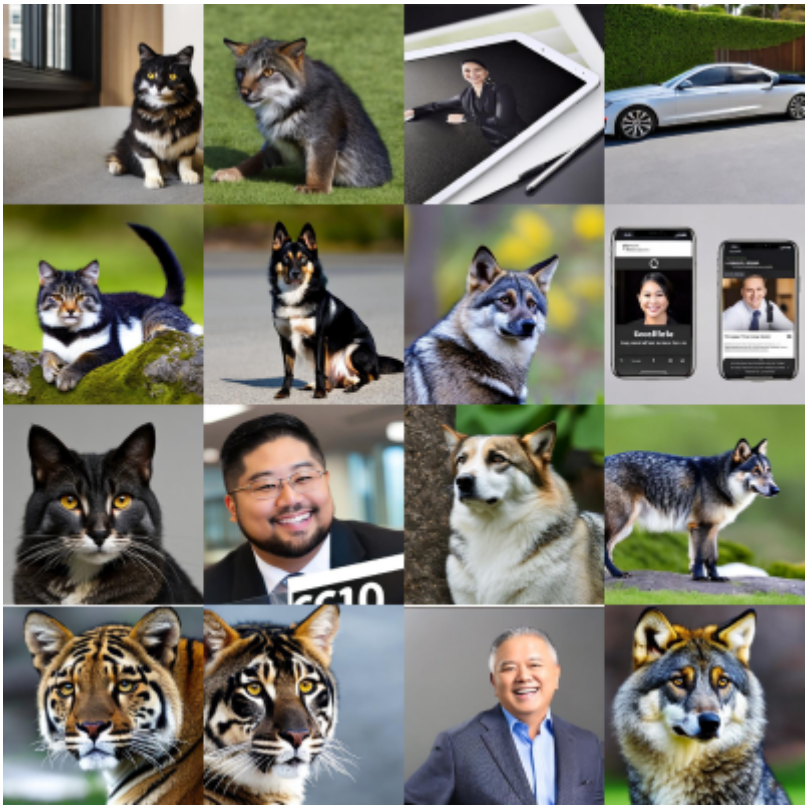

Figure 4.8: *Prompt: "Photo of a CEO" ($N = 16$), following optimisation mitigation against paired terms["he","him", "she", "her"]*

4.3.8 Embedding Degradation

In synthesising more representative embeddings, images of animals or unassociated objects were common in the image generation phase. The primary source for this degradation could be adding too many terms, pushing the embedding vector away from a boundary that indicates a term, such as "nurse". Recalling that for Stable Diffusion, the embedding dimension is $\mathbb{R}^{1024}$, the indicator is that there is noise between terms or that more careful analysis of the embedding space is required. As the dimension of a representation increases, the boundary becomes much more complex and can be considered a fragile boundary. Introducing more terms to balance against that are less related to the original term is more likely to destroy the embedding.

There are also instances of embeddings that replicate the effect of "mode collapse" where image generations concerning that concept become uniform. Figure 4.9 is built with the weights `{"engineer": 1.0, "black": -0.3268, "latino": 0.2572, "asian": -0.67352, "indian": 0.80533, "white": 0.5435}` but overwhelmingly features people with darker skin tones. In contrast, before mitigation, "engineer" also contained people in engineering attire, such as a hard hat.

Figure 4.9: *Prompt: "photo of a person who is an engineer" ($N = 16$), post mitigation, balanced against ethnicity terms directly*

We note the presence of picture frames and images that feature photographs; these come from the prompt being interpreted as a photograph. Future studies would study different settings and art styles to observe more complex interactions between a chosen scene and the occupation. Quite a few image sets also future monochrome images, or images primarily black and white, such as Figure 4.10, where the weight corresponding to the term "black" was determined to be 1. The terms "black" and "white" can be inferred to have multiple meanings, such as referring to the colour palette of the image rather than referring to a potential quantifier for the skin tone or ethnicity of a person. We opt to use this language for the experiments of this thesis primarily as BLIP is also grounded in this terminology.

The effect of setting a weight to 1 also heavily skews the distribution towards that term. Figure 4.10 is a sample from an embedding that generated only people classified as "african american". The entire distribution has shifted towards images of people with a darker skin tone, which is reflected in the finalised weights that were used to build the embedding. This indicates that the optimisation process does not penalise weights growing towards 1, even though the objective is to minimise the square of deviations, possibly in effect due to collinearity satisfying closeness for other terms. In the future, adding a regularisation term of the form $\lambda \sum_{j=1}^{k} W_j^2$ would penalise terms close to 1.

Figure 4.10: *Prompt: "photo of a person who is a lawyer" ($N = 16$), post mitigation, balanced against ethnicity terms directly*

4.3.9 Randomness of Embedding Generation

The basin-hopping process is a stochastic algorithm that randomly perturbs the coordinates to find a global minimum. The randomness of the process means that there will be a variance between each run and the final embeddings constructed. Optimising the objective is not computationally intense and involves calculating similarity scores and then adjusting, so it is affordable to generate a batch of embeddings. We examine various runs under a specified method to observe the variability between each mitigation process. Across experiments, embedding weights can vary, and a potential strategy for working with multiple runs could involve aggregating embeddings from several runs to understand how similar embeddings perform when aggregated.

Chapter 5

Conclusion

As a preliminary study, we have established an optimisation-based method for text embedding bias mitigation in text-to-image models. Our experiments show that optimising the squared deviation of similarities can induce changes in the diversity of a term. However, we also note that using a zeroth-order method like basin-hopping needs to be more precise and necessitates further examination. Future studies should examine using first-order optimisation methods with regularisation that utilise the gradient of the objective.

Our research addressed three research questions:

1. How does manipulating text embeddings affect the resulting images?

Perturbing text embeddings with weighted multiples of attribute-related terms can influence the distribution of generated images, indicating extensive potential for controlled bias mitigation at the text layer.

2. What terms can be used to balance the representation of images?

Our approach focused explicitly on single-token debiasing, where we used single-token terms such as “male”, “female”, “asian” “latino”, or “black” to balance the representation of people in images. However, these terms carry ambiguous meanings that change depending on the context, and there may be less natural terms that contribute to the diversity within a term. Future studies can focus on precisely understanding word representations within the embedding space and construct frameworks for understanding which words contribute to policy alignment for diversity.

3. How does embedding-based mitigation vary from other bias mitigation methods?

Creating debiased embeddings with our optimisation process is much faster and easier to swap out. A future study would involve comparing different bias mitigation methods for diffusion models and observing synergy in combining mitigation strategies.

To conclude, our work provides a foundational framework for mitigating bias in diffusion-based image generation models by optimising term deviations. Our zeroth order, basin-hopping process can converge to more representative embeddings, which is indicative of results for future studies on first-order optimisation processes on the text embeddings of the text encoder. Word embeddings provide a platform-agnostic method for bias mitigation that can be tested on other diffusion models, such as Stable Diffusion XL or Imagen, which use larger text encoders. Ultimately, investigating bias mitigation through word embeddings can open up debiasing opportunities on various models of varying modalities, such as text, audio, image, or video.

Bibliography

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. *Propublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bansal, H., Yin, D., Monajatipoor, M., & Chang, K.-W. (2022). How well can text-to-image generative models understand ethical natural language interventions? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1358–1370. https://doi.org/10.18653/v1/2022.emnlp-main.88

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356–4364.

Brack, M., Friedrich, F., Hintersdorf, D., Struppek, L., Schramowski, P., & Kersting, K. (2023). *SEGA: Instructing diffusion using semantic dimensions*. https://doi.org/10.48550/ARXIV.2301.12247

Cho, J., Zala, A., & Bansal, M. (2022). *DALL-eval: Probing the reasoning skills and social biases of text-to-image generative transformers*. https://arxiv.org/abs/2202.04053

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. https://doi.org/10.48550/ARXIV.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., & Kersting, K. (2023). *Fair diffusion: Instructing text-to-image generation models on fairness*. https://doi.org/10.48550/ARXIV.2302.10893

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-or, D. (2023). An image is worth one word: Personalizing text-to-image generation using textual inversion. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=NAQvF08TcyG

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). *A survey of generative AI applications*. https://doi.org/10.48550/ARXIV.2306.02781

Growcoot, M. (2023). *Controversial AI Program Generates Photorealistic Police Sketches*. https://petapixel.com/2023/02/13/controversial-ai-program-generates-photorealistic-police-sketches/ Published on PetaPixel

Heikkilä, M. (2022). *The viral AI avatar app lensa undressed me without my consent*. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/ Published on MIT Technology Review

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising diffusion probabilistic models*. https://doi.org/10.48550/ARXIV.2006.11239

Hort, M., Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2022). *Bias mitigation for machine learning classifiers: A comprehensive survey*. https://doi.org/10.48550/ARXIV.2207.07068

Joshi, B., Stewart, K., & Shapiro, D. (2017). Bringing impressionism to life with neural style transfer in come swim. *Proceedings of the ACM SIGGRAPH Digital Production Symposium*. https://doi.org/10.1145/3105692.3105697

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational bayes*. https://doi.org/10.48550/ARXIV.1312.6114

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ICML*.

Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). *Stable bias: Analyzing societal representations in diffusion models*. https://doi.org/10.48550/ARXIV.2303.11408

Nichol, A. (2022). *DALL·e 2 pre-training mitigations*. https://openai.com/research/dall-e-2-pre-training-mitigations Published on OpenAI Research

Nicoletti, L., & Bass, D. (2023). *Generative AI and bias*. https://www.bloomberg.com/graphics/2023-generative-ai-bias/ Published on Bloomberg

OpenAI. (2022). *Reducing bias and improving safety in DALL·e 2*. https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2 Announcement on the OpenAI Blog

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision*. https://doi.org/10.48550/ARXIV.2103.00020

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(1).

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with CLIP latents*. https://doi.org/10.48550/ARXIV.2204.06125

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

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., & Norouzi, M. (2022). *Photorealistic text-to-image diffusion models with deep language understanding*. https://doi.org/10.48550/ARXIV.2205.11487

Savani, Y., White, C., & Govindarajulu, N. S. (2020). Intra-processing methods for debiasing neural networks. *Advances in Neural Information Processing Systems*, *33*, 2798–2810.

Seshadri, P., Singh, S., & Elazar, Y. (2023). *The bias amplification paradox in text-to-image generation*. https://doi.org/10.48550/ARXIV.2308.00755

Silva, A. (2023). *AI images of women from around the world have gone viral. Do they promote colourism and cultural beauty standards?* https://www.abc.net.au/news/2023-08-30/artificial-intelligence-racial-bias-images-women-south-east-asia/102732046 Published on ABC News

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). *Deep unsupervised learning using nonequilibrium thermodynamics*. https://doi.org/10.48550/ARXIV.1503.03585

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following LLaMA model. In *GitHub repository*. https://github.com/tatsu-lab/stanford_alpaca; GitHub.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wiggers, K. (2023). *With firefly, adobe gets into the generative AI game*. https://techcrunch.com/2023/03/21/adobe-firefly-generative-ai/ Published on Tech Crunch

Zhang, C., Zhang, C., Zhang, M., & Kweon, I. S. (2023). *Text-to-image diffusion models in generative AI: A survey*. https://doi.org/10.48550/ARXIV.2303.07909

Appendix A

Additional Code and Data

We include any larger image datasets, code, and links not present in the paper.

A.1 Links to Repositories

The Hugging Face Textual Inversion script and notebook are available here: https://github.com/huggingface/diffusers/tree/main/examples/textual_inversion

The Embedding Inspector, which is an Automatic1111 plugin, is available here: https://github.com/tkalayci71/embedding-inspector

A.2 Image Sets

All underlying image sets that were generated for testing methods are available at: https://drive.google.com/drive/folders/1dqaLOyLD4mF-_pd3OAsIfAlR-tdn_oIs?usp=sharing

A.3 Code for training-free bias mitigation

The debiasing notebook, along with relevant scripts are available here: https://drive.google.com/drive/folders/16Xt2hcbSuPyynfyeV6PKshmM8zFwl6Oh?usp=sharing