FYP CA Report: Vision-Language Debiasing with Modality Alignment

Haoyu Zhang
National University of Singapore
e0555783@u.nus.edu

Abstract

Recently, Vision-Language (V-L) models become prevalent with widespread real-world applications. However, these powerful models have also incurred great social concerns in scenarios such as decision-making due to the inherent social biases within their image-text representations. Meanwhile, existing V-L debiasing methods still demonstrate limitions in terms of modality alignment. This research aims to fill in the gap by proposing a framework that jointly debiases both image and text embeddings in V-L models. We implement this idea by bias removal with modality alignment while preserving the original V-L representation in the models.

1 Introduction

Empowered by transformer models [40] and cross-modal pre-training objectives such as contrastive learning [48], modern V-L models (e.g., CLIP) are able to learn multi-modal representations from massive-scale image-text datasets. The pre-trained V-L models can be further fine-tuned for a wide range of downstream V-L tasks. In particular, the rich visual and linguistic concepts captured by the model during pre-training can be transferred to downstream tasks, leading to impressive model performance. Despite the promising benefits of V-L models, it is imperative to be cautious of the potential risks they can introduce to our society if they are widely applied in our daily lives. One of the most concerning risks is the social biases that V-L models sometimes exhibit in their outputs. Such social biases may arise from various factors such as biased datasets and model training, and they can severely disadvantage groups with specific attributes in decision-making scenarios. For example, a biased V-L model could potentially give predictions that unfairly decrease the chance of certain groups of people being shortlisted for an interview. Hence, mitigating the social biases hidden in the V-L models is an important research problem in the field of V-L learning.

This project focuses on debasing the pre-trained CLIP model [31] because it is the most popular V-L foudation model and has been widely used in many AI systems nowadays. Thus far, there have been efforts to measure and mitigate the social biases in CLIP. However, they are limited in their debiasing scope because they either focus on debiasing either one of the modalities in CLIP or removing image and text biases in an uncoordinated manner. Based on our exploration, both image and text bias exist in CLIP embeddings and they manifest differently, each with its own distinct characteristics and implications. As a result, existing methods which adopt unaligned debiasing in both modalities may lead to suboptimal debiasing outcomes and even undermine the quality of the original V-L representation in CLIP. To address this critical problem, this research project aims to explore the image and text bias in CLIP and propose a method that coordinates both image and text debiasing to achieve more effective debiasing with minimum impact on the original representation learning capability of CLIP.

For acknowledgement, this research is conducted in collaboration with Dr. Yangyang Guo and under the supervision of Prof. Mohan Kankanhalli.

Preprint. Under review.

2 Literature Review

2.1 Visual-Language Pre-Trained Models (VL-PTMs)

Several types of deep learning models such as RNN [2] and CNN [19] have been proposed for V-L learning. However, they are task-specific and can hardly be transferred to other V-L tasks. With the development of transformer models [40] in recent years, transformer-based VL-PTMs have grown in popularity due to their superior performance and transferability.

Most existing VL-PTMs can be categorized into two types, single-stream and dual-stream, according to their architectures [13]. Single-stream VL-PTMs such as ViLBERT [23] and ViLT [20] use a single transformer encoder to model both intra-modal and inter-modal interaction. On the other hand, dual-stream VL-PTMs such as CLIP [31] and ALIGN [17] leverage two separate single-modality transformer encoders to model intra-modal interaction in image embeddings and text embeddings respectively. Subsequently, techniques such as multi-modal contrastive learning [48] are used to model the cross-modal interaction between image embeddings and text embeddings. Some models such as ALBEF [21] and LXMERT [39] also introduce an additional cross-modality encoder to model complex image-text interaction.

After pre-training, VL-PTMs can be fine-tuned on various V-L tasks such as Visual Question Answering [3], image-text retrieval [7] and Visual Entailment [46]. Some VL-PTMs such as CLIP also have remarkable zero-shot performance for classification and retrieval tasks [31]. Additionally, because of the alignment of the pre-trained text and image encoder, VL-PTMs are also used in image-to-text generative models. Stable Diffusion [34], for example, incorporates a pre-trained CLIP text encoder for image generation conditioned on the text embeddings of CLIP.

2.2 Bias in Natural Language Processing (NLP)

Existing work focuses on social biases with respect to common sensitive topics such as gender, race and religion in language modelling. These biases mainly come from problematic text corpora which are unbalanced or contain stereotypical language. Most early work in NLP bias studies social bias in word-level embeddings and there are several metrics such as the World Embedding Association Test (WEAT) [6] and projection on gender direction [5] to measure social biases in word embeddings. Projection-based methods such as Hard-Debiasing [5] is proposed to identify a bias subspace and debias word embeddings by removing biases projected onto the bias subspace.

More recently, with the development of pre-trained language models such as BERT [11] and GPT [30], sentence-level embeddings are widely used and more work has started to focus on sentence-level social biases. As an extension of WEAT, Sentence Encoder Association Test (SEAT) [24] is proposed to measure the bias in sentence embeddings, followed by additional bias benchmarks such as SteroSet [26] and CrowS-Pairs [27]. Inspired by projection-based word-embedding debiasing, methods such as SentenceDebias [22] and Iterative Nullspace Projection [32] are developed to remove biases through vector-space manipulation. Other approaches based on prompting [35], dropout regularization [45] or contrastive learning [8] have also been proposed to debias pre-trained text encoders.

2.3 Bias in Computer Vision (CV)

In the context of CV, common social biases such as gender and racial bias have also been extensively studied. Most work focuses on CV tasks such as classification and the social biases are measured in terms of task-specific metrics such as Demographic Parity [14] and Equalized Odds [16]. These metrics evaluate whether the performance of CV models varies significantly when applied to different social groups. Several distinct approaches are proposed to mitigate bias in CV models by adding confusion loss during training [1], leveraging adversarial training to directly learn the debiased representation of images [44], learning a fair module [12] to remove the bias, or training the model by incorporating additional explanations to enhance robustness against undesired correlations [37]. These techniques are also used beyond the scope of social bias to address more general biases caused by spurious correlations [33].

2.4 Bias in CLIP

Due to the multi-modal nature of CLIP, existing studies on CLIP bias are heavily influenced by work from both NLP and CV, targeting social bias regarding gender, race and age in CLIP. A variety of the bias metrics are used to evaluate the bias in CLIP. Some work adopts NLP metrics such as WEAT to measure the preferential bias in image and text embeddings of CLIP [4]. Some other studies further extend classification-based metrics in CV to the multi-modal domain and propose a retrieval-based bias metric to evaluate fairness based on the proportion of images retrieved with different groups [36]. Since the CLIP text encoder is used in text-to-image generative models such as Stable Diffusion [34], some research also evaluates the bias in CLIP by measuring the diversity of the images generated conditioned on CLIP text embeddings [9].

Existing CLIP debiasing methods can be divided into two categories: embedding vector manipulation and fair module learning. Methods in the former category are training-free and directly remove the bias information from image or text embeddings through bias subspace projection [9] or feature pruning [42]. On the other hand, methods based on a fair module use fairness datasets to train a learnable module that can be added in CLIP to remove bias produced by CLIP. The learnable module can be trained with adversarial learning [4] or other optimization goals [43] to remove irrelevant correlations from the CLIP embeddings. Regarding debiasing scope, most debiasing approaches focus on single-modal debiasing of image or text embeddings in CLIP. Only [42] attempts multi-modal debiasing by removing certain image and text dimensions from CLIP embeddings. However, the removal of dimensions in a coarse-grained manner leads to significant information loss.

Since existing CLIP debiasing methods fail to achieve the joint removal of image and text bias with modality alignment, the biases in CLIP may not be fully removed in both image and text embeddings, leading to sustained social biases in V-L tasks. Additionally, the alignment of image and text embeddings in CLIP may also be compromised due to uncoordinated image and text debiasing, leading to degradation of V-L task performance.

2.5 Fairness Datasets

Different from NLP debiasing methods which can swap keywords in sentences to create biased text prompts, CV debiasing requires annotated images with different social biases. Several fairness datasets have been curated for training fair CV models, such as FairFace [18] and UTKFaces [49], which have labelled face images from different age, race and gender groups. Recently, a more comprehensive dataset, FACET [15], has been proposed, providing full-body human images with a greater variety of labels such as occupation and skin tone apart from typical gender and age information.

3 Research Methodology

3.1 Problem Statement

Debiasing CLIP is essentially the problem of producing unbiased joint text-image representations. To define the idea of fairness, consider a set of protected attributes, denoted as $A = \{a_1, \ldots, a_m\}$, and a set of target concepts, denoted as $C = \{c_1, \ldots, c_n\}$. Protected attributes refer to innate attributes of a person such as gender, race and age. Target concepts refer to specific properties that a person can have, such as occupations and personal qualities. If a model is fair, it assigns the target concepts independent of the protected attributes. In other words, the attributes are protected from being taken into consideration by the model when making a decision. For example, an AI model used for resume shortlisting can be considered fair if it decides to shortlist (target concept) a candidate regardless of gender (protected attribute).

In the context of CLIP, the model takes in an image-text pair, denoted as (I_a, T_c) where I_a is an image and T_c is a sentence. Suppose I_a has a protected attribute $a \in A$ and T_c has a target concept $c \in C$, in order to be fair, the matching of (I_a, T_c) should be independent of the protected attribute a of I_a. This means that if a is switched to a different $a' \in A$ in I_a, the matching score of $(I_{a'}, T_c)$ should be close to that of (I_a, T_c). For instance, for a fair CLIP model, the matching score of an image of a male engineer, and the sentence "This is a photo of an engineer." should be similar to the matching score of an image of a female engineer and the same sentence "This is a photo of an

engineer.", assuming that the female and male engineer images are similar. A more rigorous way to measure the fairness (bias) of the CLIP model will be covered in Section 3.5.

Since CLIP has an image encoder E_i and a text encoder E_i, and the similarity score of any (I, T) pair is calculated by:

$$\text{Sim}(I, T) = \frac{E_i(I) \cdot E_t(T)}{\|E_i(I)\| \|E_t(T)\|},$$

both image or text bias in E_i or E_t can affect the overall matching score, leading to bias in CLIP. For instance, the image bias reflected by a major difference in $E_i(I_a)$ and $E_i(I_{a'})$ will affect the matching of (I_a, T_c) and $(I_{a'}, T_c)$. Hence, to obtain a fair CLIP model, we can develop a debiased image encoder $E_i{}'$ such that $E_i{}'(I_a)$ is close to $E_i{}'(I_{a'})$. In parallel, any text bias in $E_t(T_c)$ may also result in a change in the matching score if $E_t(T_c)$ is biased towards $E_i(I_{a'})$ or $E_i(I_a)$. We can also develop a debiased text encoder $E_t{}'$ to address the bias. Hence, the debiasing of CLIP can be broken down into the debiasing of image and text encoders. Specifically, both debiased $E_i{}'$ and $E_t{}'$ can be developed to ensure that the matching of image and text is independent of protected attributes. How to debias image and text encoder to achieve the fairest CLIP still remains an open problem, which will be addressed in this research project.

3.2 Research Objectives

Given the limitations of existing CLIP debiasing methods as discussed in the literature review, the ultimate goal of this research project is to develop a debiasing framework which achieves the modality alignment in image and text debiasing in CLIP. We expect that the joint image and text debasing with modality alignment will enhance the fairness of the CLIP model in both discriminative and generative tasks. To accomplish this goal, preliminary experiments have been conducted to examine the manifestation of image and text bias in CLIP embeddings. Based on the understanding of specific patterns of image and text bias, a novel debiasing method will be designed to debias both image and text embeddings with modality alignment such that the V-L bias can be removed more effectively with less drop in V-L performance of the debiased CLIP.

3.3 Exploration on Image and Text Bias in CLIP

In order to align the modalities for image and text debiasing, we have conducted preliminary experiments to investigate the image and text bias in CLIP. Specifically, a qualitative analysis of image and text bias in CLIP is first performed. Subsequently, the alignment of image and text bias is evaluated. Finally, an additional test is conducted to explore the biases in CLIP arising from cross-modal interaction.

The experiments are conducted using handcrafted text prompts and images with different occupations and genders from the FACET dataset. Only gender bias has been investigated in our experiments so far, and other biases related to race or age will be evaluated in the next stage.

3.3.1 Visualisation of Image and Text Bias

To obtain an intuitive understanding of the bias in the text encoder and image encoder of CLIP, we have visualised the gender biases using t-SNE plotting. For texts, we use a prompt sentence template: "This is a photo of a {male/female} {OCCUPATION}." Following the set-up of FACET, 52 occupation keywords such as "dancer", "lawman" or "astronaut" are filled in the template, and by swapping the gender (male or female), a total number of 104 sentences with different combinations of occupations and genders are produced. The text encoder of the pre-trained CLIP is used to encode these 104 sentences to generate 104 text embedding vectors with a dimension of 512. Similarly, for images, 104 images are randomly selected from FACET with different combinations of occupation and gender corresponding to the sentences. The 104 images are fed into the image encoder of CLIP to produce 104 image embedding vectors also with a dimension of 512.

To investigate the distributions of the biased image and text embeddings, we use t-SNE to visualise the high-dimensional embeddings in a 2-D plot. To differentiate genders and modalities, four classes are introduced: male+text, male+image, female+text and female+image. Each class has 52 points, corresponding to 52 occupations with the gender and modality corresponding to the class. It can be observed from Figure 1 that the text embeddings and image embeddings form two different clusters.

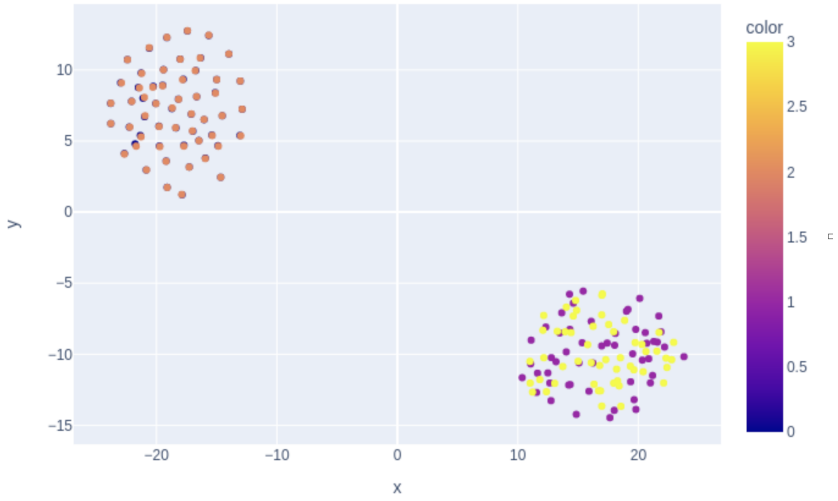

Figure 1: The t-SNE plot of image and text embeddings with different genders. Color 0: male+text; color 1: male+image; color 2: female+text; color 3: female+image.

The 104 text embeddings with male and female genders are all located in the upper-left corner, whereas the 104 image embeddings with male and female genders are located in the lower-right corner. Within the text embedding cluster, there are 52 smaller pairs with almost overlapping points. Each pair corresponds to an occupation with male and female genders. The overlapping of most pairs with the same occupation and different genders suggests that gender bias is not significant in the text encoder because the oppositely biased sentences in each pair have similar text embeddings.

On the other hand, in the image embedding cluster, there are no obviously aligned pairs of points, suggesting that the embeddings of images with the same occupation but different genders are not very similar compared with the text embeddings. This indicates that the biases in image embeddings are possibly more significant. However, the difference between image and text bias cannot be quantified from visualisation because the t-SNE plot does not necessarily preserve accurate distance information. Hence, this visualisation only suggests that the image and text bias may not be aligned. Further investigation is required to assess the alignment of both biases.

3.3.2 Alignment of Image and Text Bias Subspaces

To investigate how different the distributions of image and text biases are from each other, we have designed an experiment to further determine the alignment of image and text bias subspaces.

For any image and text embedding, $E_i(I)$ or $E_t(T)$, we follow the analysis in [36] to split the embedding into two additive components: the gender bias information ($\phi_i(I)$ or $\phi_t(T)$) and the gender-neutral information ($\overline{\phi}_i(I)$ or $\overline{\phi}_t(T)$). This can be denoted as:

$$E_i(I) = \overline{\phi}_i(I) + \phi_i(I) \tag{1}$$

$$E_t(T) = \overline{\phi}_t(T) + \phi_t(T) \tag{2}$$

Suppose there are two images with opposite gender attributes (male and female) and the same concept (e.g. teacher), denoted as I_m and I_f, and there are also two sentences with opposite gender keywords (male and female) and the same concept (e.g. teacher), denoted as T_m and T_f. Substituting them into the equation 1 and 2 gives:

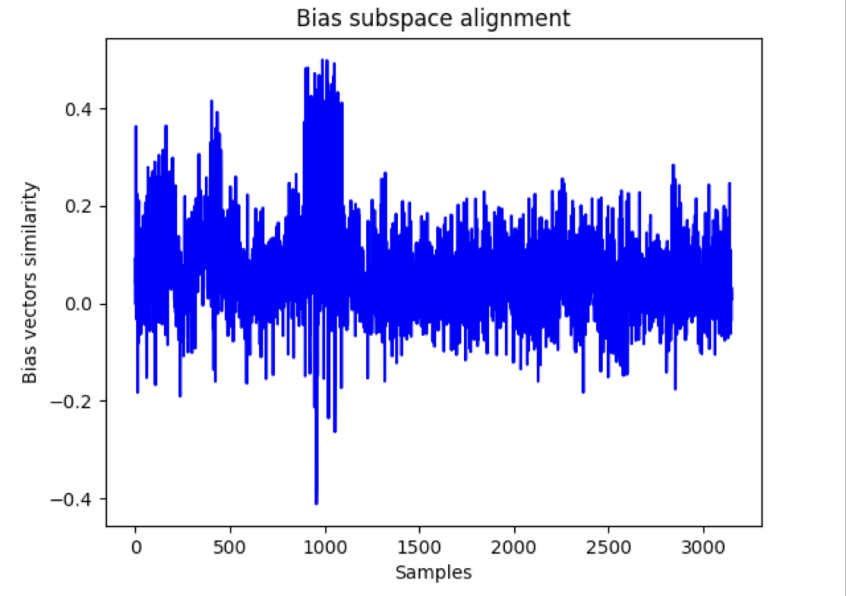

Figure 2: The cosine similarities between the bias difference vectors from image and text bias subspaces across all samples

$$E_i(I_m) = \overline{\phi}_i(I_m) + \phi_i(I_m) \tag{3}$$
$$E_i(I_f) = \overline{\phi}_i(I_f) + \phi_i(I_f) \tag{4}$$
$$\Rightarrow E_i(I_m) - E_i(I_f) = \phi_i(I_m) - \phi_i(I_f) \tag{5}$$
$$E_t(T_m) = \overline{\phi}_t(T_m) + \phi_t(T_m) \tag{6}$$
$$E_t(T_f) = \overline{\phi}_t(T_f) + \phi_t(T_f) \tag{7}$$
$$\Rightarrow E_t(T_m) - E_t(T_f) = \phi_t(T_m) - \phi_t(T_f) \tag{8}$$

Equation 5 and 8 are obtained assuming the gender-neutral information is approximately the same and can be cancelled off for texts and images with the same occupation but different genders.

Assuming the image and text biases are distributed in a similar manner, their bias subspace should be aligned. For example, for an image-text pair (I_m, T_m) with the same concept (e.g. teacher) and the gender of male, $\phi_i(I_m)$ and $\phi_t(T_m)$ should have aligned subspaces. Similarly, for an image-text pair (I_f, T_f) with the same concept as (I_m, T_m) (e.g. teacher) but the gender of female, $\phi_i(I_f)$ and $\phi_t(T_f)$ should also have aligned subspaces. Additionally, $\phi_i(I_m)$ and $\phi_i(I_f)$ share the same image bias subspaces, and $\phi_t(T_m)$ and $\phi_t(T_f)$ share the same text bias subspaces. Based on the assumption that image and text bias are aligned, the "image bias difference vector" $\phi_i(I_m) - \phi_i(I_f)$ should also align with the "text bias difference vector" $\phi_t(T_m) - \phi_t(T_f)$. Hence, the alignment of the image and text bias subspaces can be evaluated by the alignment of $\phi_i(I_m) - \phi_i(I_f)$ and $\phi_t(T_m) - \phi_t(T_f)$ for each sample that contains two image-text pairs, (I_m, T_m) and (I_f, T_f) which all correspond to the same concept.

According to the equation 5 and 8, $\phi_i(I_m) - \phi_i(I_f)$ and $\phi_t(T_m) - \phi_t(T_f)$ can be calculated with the embeddings of (I_m, T_m) and (I_f, T_f). Hence, we sampled ~3000 samples of (I_m, T_m) and (I_f, T_f) from the FACET dataset. Samples with various occupations are selected to ensure variety. For each sample, $\phi_i(I_m) - \phi_i(I_f)$ and $\phi_t(T_m) - \phi_t(T_f)$ are calculated and their alignment is evaluated based on cosine similarity. If the image and text bias subspaces align well with each other, the alignment between the bias difference vectors for each sample should be roughly the same. However, after calculating the bias alignment scores for each sample and comparing them across samples through plotting, the alignment scores vary significantly across the ~3000 samples, according to Figure 2. This suggests that the bias subspaces may not align with each other. Additionally, we also consider the possibility that the bias subspaces may have similar shapes but different orientations. Hence, we

have explored training a linear transformation to transform the bias difference vector from one bias subspace to the other. However, after trying with linear regression, the R^2 value is as low as 0.282, suggesting that even the shapes of the bias subspaces may not match.

3.3.3 Bias from Cross-Modal Interaction

Previous experiments have assessed biases in either text or image modality. However, they cannot completely reflect the full picture of biases in the CLIP embeddings because the image-text interaction is not considered. Hence, a test is also conducted to investigate the biases from a cross-modal interaction perspective.

In CLIP, the bias can be reflected by the change in matching between image and text due to gender association. Since images can be matched to texts or vice versa, we conduct tests in both ways. An illustration of the tests is shown in Figure 3.

Matching an image to text prompts This test consists of two stages. At the first stage, we choose an image with a specific gender and concept. We match this image to a range of sentences with different concepts as "text prompts". For example, we can choose an image of a male teacher. We match this image to sentences with the template "A photo of a {CONCEPT}.", and we substitute concept keywords such as teacher, athletes or other occupations. The image should match to the sentence "A photo of a teacher." with the highest cosine similarity because they have the same concept.

In the second stage, in order to elicit gender bias, we add a description of the gender information in the sentence template, where the gender information matches the gender of the image. For example, now we match an image of a male teacher to sentences with the template "A photo of a *male* {CONCEPT}." and substitute concept words into the template. If the image is still matched to the corresponding sentence "A photo of a male teacher" with the highest similarity, then the matching is not affected by additional gender information. However, if the image is matched to another sentence "A photo of a male farmer", this may suggest that a stronger gender association leads to the change of matching, indicating bias in CLIP. Specifically, the correct concept "teacher" has been overridden due to bias.

For the first-stage test, we conduct 416 rounds of matching of an image to 52 text prompts. Images of 52 different occupation-related concepts such as "teacher", "dancer" and "lawman" are taken from the FACET dataset, and for each concept we randomly choose 4 male and 4 female images, and 416 images are used overall. For each of the 416 images, we generate its image embedding using a pre-trained CLIP image encoder and calculate the cosine similarity between this image embedding and text embeddings of 52 different sentence prompts with different concepts. The sentence prompt with the highest similarity to the image is chosen as the final match. Among the 416 rounds of matching, there are 219 rounds in which the image is finally matched to a sentence prompt with the same concept. The correct concept matching rate is 52.6%.

For the second-stage test, we conduct another 416 rounds of matching with the same images but gender-biased sentence prompts. Among the 416 rounds of matching, there are also 219 rounds of the same concept matching. The correct concept matching rate remains 52.6%. This shows that increasing the gender association between the image and text prompts may not lead to more cases of overriding of the original concept.

Matching a text to image prompts This test is also conducted in two stages in a similar way. In the first stage, we choose a sentence with a specific concept. We match this sentence to a range of images with different concepts and the same gender as "image prompts". For example, we can choose the sentence "A photo of a teacher." and match this sentence to images of a male teacher or males of other occupations. The sentence should be matched to the image of a male teacher with the highest cosine similarity because they have the same concept.

In the second stage, in order to elicit gender bias, we add a description of the gender information in the sentence, where the gender information matches the gender of the images. For example, now we match the sentence "A photo of a *male* teacher" to the images of a male teacher or males of other occupations. If in this round the sentence is still matched to the corresponding image of a male teacher with the highest similarity, then the matching is not affected by additional gender information. However, if the sentence is matched to another image of a male farmer, this may suggest that a stronger gender association leads to the change of matching, indicating bias in CLIP.

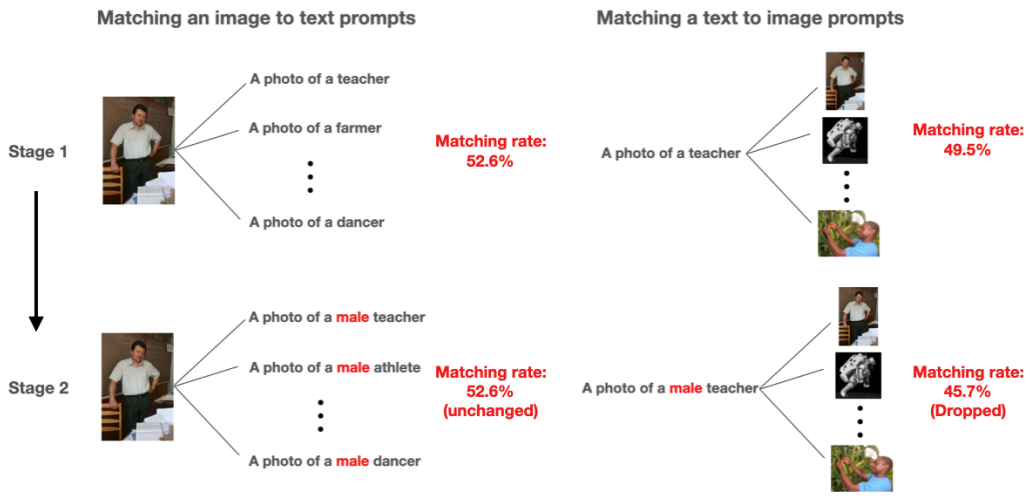

Figure 3: Illustration of the image-text and text-image matching tests

For the first-stage test, we have conducted 416 rounds of matching of a text to 52 image prompts. There are 52 sentences with different concepts. For each round, one sentence is matched to 52 images of different occupation-related concepts such as "teacher", "dancer" and "lawman". The image with the highest similarity to the sentence is chosen as the final match. We vary the images for more rounds of matching. For each occupation, we choose 4 males and 4 females, and 416 images are used in total. Among the 416 rounds of matching, there are 206 rounds in which the sentence is finally matched to an image prompt with the same concept. The correct concept matching rate is 49.5%.

For the second-stage test, we conduct another 416 rounds of matching with the gender-added sentence and the same image prompts. Among the 416 rounds of matching, there are only 190 rounds of the same concept matching. The correct concept matching rate drops to 45.7%. This suggests that an increased level of gender association between the text and its image prompts leads to more significant overriding of the original concept.

Analysis of results It is found that a stronger gender association between the text and its image prompts leads to observable biases which are reflected by the more serious overriding of the correct concepts due to gender association. We think that these biases indicate that in the image modality, some concepts are biased towards certain gender attributes, and a stronger gender association between the text prompt and the image may lead to a stronger association between the text prompt and the biased concept, resulting in the overriding. On the other hand, a stronger gender association between the image and its text prompts does not increase the overriding of correct concepts, suggesting that in the text modality, text-based concepts are less biased towards gender attributes compared to that of the image modality. This supports the observation based on the t-SNE visualisation of image and text bias.

3.3.4 Comments on Bias in CLIP

Based on the analysis of the biases exhibited in text and image embeddings, it is found that the gender biases in image embeddings and text embeddings are not aligned but rather distributed very differently from each other. Additionally, compared with text embeddings, there are more significant biases in image embeddings in the form that some concepts may have a stronger association with gender attributes. Consequently, when debiasing image and text encoders, we should be aware of the different manifestations of image and text biases.

In the next stage of our research, the bias in CLIP will be further investigated with extended scope which focuses on other social biases such as race and age. Additionally, existing preliminary experiments will be conducted again with more samples to further verify the discoveries.

3.4 Proposed Method

Based on our discovery that image and text bias are distributed differently, we have formulated a general idea to jointly debias the image and text embeddings. Firstly, an alignment module will be trained to transform the bias in image embeddings to the same bias subspace as text embeddings. After the alignment, a universal debiasing method will be applied to the transformed image embeddings and text embeddings to remove their biases simultaneously. However, due to time constraints, our ideas are not finalised and are still subject to changes.

3.4.1 Alignment of Text and Image Bias

Firstly, the image bis will be aligned with the text bias through an alignment module that transforms the image bias into the same bias space as the text bias. The reason for aligning image bias with text bias instead of the other way around is based on the observation that biased text prompts can be easily created by switching biased keywords. On the other hand, designing biased image prompts is much more challenging. Inspired by existing debiasing work in NLP, we plan to use biased text prompts to elicit and remove the bias in text embeddings and the corresponding image embeddings. Aligning the image bias with text bias beforehand can ensure a more thorough removal of text and image bias.

A single-layer neural network is tentatively chosen as the alignment module. The rationale for using a module with only one hidden layer is to reduce the number of parameters to ease the learning of the alignment module. However, if the real performance does not meet our expectations, a deeper neural network with more modelling power will be used to learn the alignment. Thus far, how to supervise the training of the alignment module is still under consideration.

3.4.2 Debiasing of Text and Image Embeddings

For joint text and image debiasing, we tentatively plan to adopt the debiasing method proposed by FairFil [8] to train a neural fair module through contrastive learning. In the original FairFil work, the fair module trained with biased text prompts is able to remove complex biases in text embeddings. We plan to modify the FairFil framework to involve both images and texts in the contrastive learning process to achieve image-text debiasing. However, the exact design is still under consideration. We may also adopt other methods depending on our experiment results.

For text debiasing, we plan to design biased text prompts by editing the corpora from WikiText-2 [25], Stanford Sentiment Treebank [38], Reddit [41], MELD [29] and POM [28], following the set-up of Fairfil [8]. For image debiasing, we will use labelled images from UTKFace [49], FairFace [18] and FACET[15].

3.5 Evaluation Metrics

The evaluation of a CLIP debiasing method consists of two criteria: bias and performance of the debiased CLIP. A method meeting both criteria will ensure the fairness and usability of CLIP in real-world applications.

To measure the bias, we will follow the retrieval-based metric, the mean $MaxSkew@k$, which is used in various CLIP debiasing papers [4; 36]. This metric indicates the largest unfair advantage given to images with a certain attribute in the retrieval task. To calculate the mean $MaxSkew@k$, we first calculate the $MaxSkew@k$ for a specific text prompt t, denoted as $MaxSkew@k(t)$. The text prompt t has a neutral concept (e.g. "teacher"), and it is used to retrieve k most similar images from a pool of images. Each image has a protected attribute, $a \in A$, where A is a set of protected attributes such as genders, races, and ages. The $MaxSkew@k(t)$ can be calculated as:

$$MaxSkew@k(t) = \max_{a \in A} \ln \left(\frac{p_{t,a}}{p_a} \right)$$

where $p_{t,a}$ denotes the proportion of images with attribute a in the retrieved k images, and p_a denotes the proportion of images with attribute a in the original pool of images. $MaxSkew@k(t)$ having a value of 0 indicates absolute fairness regarding the concept in t because it suggests that the proportion of different groups in the retrieved images stays the same compared to the original pool, meaning that no groups of images with specific attributes are under-represented or over-represented in the retrieved sets of images.

The mean $MaxSkew@k$ can be further calculated by using different $t \in T$ where T is the set of all text prompts with a wide range of neutral concepts. A mean $MaxSkew@k$ close to 0 indicates the fairness of the debiased CLIP model regarding all tested neutral concepts in T.

To measure the V-L performance of the debiased CLIP model, we will evaluate its zero-shot performances in classification tasks (ImageNet [10], etc.) and image retrieval tasks (flickr-1k [47], etc.) with standard metrics such as accuracy and R@5. Additionally, we also plan to apply the debiased CLIP text encoder to Stable Diffusion to evaluate whether our method has improved the fairness in CLIP-driven generative models. Particularly, the fairness of generative models can be evaluated by quantitative metrics such as the balance between males and females among all the images generated. Additionally, qualitative observations such as the diversity of attributes in the generated images can also be used to assess the fairness of generative models.

4 Research Plan

The general timeline for the research project is proposed to keep track of the progress.

Time Period	Expected Progress
Aug 2023 | Oct 2023	Literature review; conduct preliminary experiments
Nov 2023 | Dec 2023	Design and implement debiasing method focusing on gender bias
Jan 2024 | Feb 2024	Training; evaluate results; extend the scope to other social biases
Mar 2024 | April 2024	Refine the method; prepare the final report and presentation

References

[1] M. Alvi, A. Zisserman, and C. Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Sep 2018.

[2] A. Anastasopoulos, S. Kumar, and H. Liao. Neural language modeling with visual features, 2019. URL `https://arxiv.org/abs/1903.02930`.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

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

[5] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Cornell University - arXiv,Cornell University - arXiv*, Jul 2016.

[6] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, page 183–186, Apr 2017. doi: 10.1126/science.aal4230. URL `http://dx.doi.org/10.1126/science.aal4230`.

[7] M. Cao, S. Li, J. Li, L. Nie, and M. Zhang. Image-text retrieval: A survey on recent research and development, 2022. URL `https://arxiv.org/abs/2203.14713`.

[8] P. Cheng, W. Hao, S. Yuan, S. Si, and L. Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders. *Learning,Learning*, Mar 2021.

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

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2009. doi: 10.1109/cvpr.2009.5206848. URL http://dx.doi.org/10.1109/cvpr.2009.5206848.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

[12] M. Du, S. Mukherjee, G. Wang, R. Tang, A. Awadallah, and X. Hu. Fairness via representation neutralization. *Neural Information Processing Systems,Neural Information Processing Systems*, Dec 2021.

[13] Y. Du, Z. Liu, J. Li, and W. X. Zhao. A survey of vision-language pre-trained models, 2022. URL https://arxiv.org/abs/2202.10936.

[14] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. *arXiv: Machine Learning,arXiv: Machine Learning*, Dec 2014.

[15] L. Gustafson, C. Rolland, N. Ravi, Q. Duval, A. Adcock, C.-Y. Fu, M. Hall, and C. Ross. Facet: Fairness in computer vision evaluation benchmark. Aug 2023.

[16] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *arXiv: Learning,arXiv: Learning*, Oct 2016.

[17] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[18] K. Karkkainen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan 2021. doi: 10.1109/wacv48630.2021.00159. URL http://dx.doi.org/10.1109/wacv48630.2021.00159.

[19] A. Khamparia, B. Pandey, S. Tiwari, D. Gupta, A. Khanna, and J. Rodrigues. An integrated hybrid cnn–rnn model for visual description and generation of captions. *Circuits, Systems, and Signal Processing*, 39, 02 2020. doi: 10.1007/s00034-019-01306-8.

[20] W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021. URL https://arxiv.org/abs/2102.03334.

[21] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi. Align before fuse: Vision and language representation learning with momentum distillation, 2021. URL https://arxiv.org/abs/2107.07651.

[22] P. Liang, I. Li, E. Zheng, Y. Lim, R. Salakhutdinov, and L.-P. Morency. Towards debiasing sentence representations. *Cornell University - arXiv,Cornell University - arXiv*, Jul 2020.

[23] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. URL https://arxiv.org/abs/1908.02265.

[24] C. May, A. Wang, S. Bordia, S. Bowman, R. Rudinger, A. Adam, N. Harry, A. Ellen, and P. Paul. On measuring social biases in sentence encoders.

[25] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models, 2016.

[26] M. Nadeem, A. Bethke, and S. Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv: Computation and Language,arXiv: Computation and Language*, Apr 2020.

[27] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Jan 2020. doi: 10.18653/v1/2020.emnlp-main.154. URL http://dx.doi.org/10.18653/v1/2020.emnlp-main.154.

[28] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency. Computational analysis of persuasiveness in social multimedia. In *Proceedings of the 16th International Conference on Multimodal Interaction*, Nov 2014. doi: 10.1145/2663204.2663260. URL http://dx.doi.org/10.1145/2663204.2663260.

[29] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jan 2019. doi: 10.18653/v1/p19-1050. URL http://dx.doi.org/10.18653/v1/p19-1050.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners.

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

[32] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection, 2020.

[33] L. Rieger, C. Singh, W. Murdoch, and B. Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. *arXiv: Learning,arXiv: Learning*, Sep 2019.

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

[35] T. Schick, S. Udupa, and H. Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, page 1408–1424, Dec 2021. doi: 10.1162/tacl_a_00434. URL http://dx.doi.org/10.1162/tacl_a_00434.

[36] A. Seth, M. Hemani, and C. Agarwal. Dear: Debiasing vision-language models with additive residuals. Mar 2023.

[37] K. Singh, D. Mahajan, K. Grauman, Y. Lee, M. Feiszli, and D. Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. *Cornell University - arXiv,Cornell University - arXiv*, Jan 2020.

[38] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, and S. Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, Oct. 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

[39] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers, 2019. URL https://arxiv.org/abs/1908.07490.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

[41] M. Völske, M. Potthast, S. Syed, and B. Stein. TL;DR: Mining Reddit to learn automatic summarization. In L. Wang, J. C. K. Cheung, G. Carenini, and F. Liu, editors, *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL https://aclanthology.org/W17-4508.

[42] J. Wang, Y. Liu, and X. Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Sep 2021.

[43] J. Wang, Y. Zhang, and J. Sang. Fairclip: Social bias elimination based on attribute prototype learning and representation neutralization. Oct 2022.

[44] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *Cornell University - arXiv,Cornell University - arXiv*, Nov 2018.

[45] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, and S. Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv: Computation and Language,arXiv: Computation and Language*, Oct 2020.

[46] N. Xie, F. Lai, D. Doran, and A. Kadav. Visual entailment: A novel task for fine-grained image understanding, 2019.

[47] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, page 67–78, Dec 2014. doi: 10.1162/tacl_a_00166. URL `http://dx.doi.org/10.1162/tacl_a_00166`.

[48] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *CoRR*, abs/2010.00747, 2020. URL `https://arxiv.org/abs/2010.00747`.

[49] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder.