Bias Across Different Modalities in VLMs: Evaluation-2

Vysishtya Karanam
2022102044

Kriti Madumadukala
2022101069

Jahnavi Venkamsetty
2022101118

Vivek Hruday Kavuri
2022114012

Abstract

Frozen Problem Statement: We aim to determine whether text or vision contributes more to gender bias in Vision-Language Models (VLMs). To achieve this, we will apply two debiasing techniques—Counterfactual Data Augmentation (CDA) and Task Vector—separately to each modality. Our hypothesis is that the modality showing the least bias after debiasing was originally the most biased. For our experiments, we will use CLIP and PaliGemma-2 as the models and the CelebA-Dialogue dataset, which we will annotate for gender and stereotypes.

1 Introduction

For this phase of evaluation, we experiment using the CLIP model with the method Counterfactual Data Augmentation (CDA). We use Resolution accuracies (Male, Female, Average) and Gender Gap (GG) as our evaluation metrics. We use the test split of the same dataset that we use for debiasing to evaluate the bias. Our results till this evaluation tell us that the text modality contains greater bias than the vision modality.

2 Methodology

2.1 Data Curation

We use the CelebA-Dialogue dataset and curate the samples from the same. This dataset contains structured annotations describing different facial attributes of celebrities and rating of each of the attributes on a scale of 0 to 5. The captions also include gender specific pronouns such as "she" , "her", "he", "him" etc.. indicating the possibility of an implicit gender labeling task.

2.2 Data Pre-processing and Annotation

First of all, we need labels of gender for every data point. So, we used a rule-based automatic labeler. We search for gender related terms or pronouns such as "his/her", "he/she", "gentleman, lady", "male/female" etc... and based on the presence of these words, we label the data point as either male or female. We assume that annotating this way would work because, we have set the rule based annotator such that if none of these words are present, we ask it to label the data point as "unknown", which only give us 40 data points with "unknown" label. This is negligible compared to the total size of the dataset and hence we prune them.

Then, we annotate the data points for stereotype. The dataset come with a rating from 0-5 for each of the data point on each of attributes {Bangs, Smiling, No Beard, Young, Eye glasses}. Based on the

Table 1: CDA on Anti-stereotypical male data

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Raw CLIP	0.5798	0.9856	0.7827	0.4058
Vision Frozen	0.9789	0.9892	0.9840	0.0103
Text Frozen	0.7535	0.9757	0.8646	0.2222
None Frozen	0.9859	0.9811	0.9835	0.0048
Everything except Projections Frozen	0.6479	0.9811	0.8145	0.3332

rating of a data point and the threshold we set for stereotypical male and female, we annotate the data point as either a stereotypical example or an anti-stereotypical example. Our thresholds are set by referring to a couple publications and the statistics from the data too support the thresholds. For the details of the publications or the plots from the data, refer appendix.

2.3 Model Debiasing

For this phase, the model we use is *clip-vit-base-patch32*. Then, to augment counterfactual data, we use the data we annotate we describe in the previous sub-section. We use Resolution Accuracy (Male, Female and Avg.) and Gender Gap as our evaluation metrics. We use the data with gender and stereotype attributes as "male" and "true" since we have noticed a bias towards female gender. But then, we also use both genders with stereotypical attribute as "true" in another setting because of our results increasing sensitivity towards male. We show the results related to the former one in the coming sections and analyze them but the results for the latter setting can be found in the appendix.

3 Evaluation

To quantify gender bias in Vision-Language Models (VLMs), we employ **Resolution Accuracy (RA)** as our primary metric. **RA** measures classification performance for male (RA_m) and female (RA_f) labels by evaluating how accurately the model assigns gendered labels to images.

We define the **Average Resolution Accuracy** (RA_{avg}) as the mean accuracy across male and female classifications:

$$RA_{avg} = \frac{RA_m + RA_f}{2} \tag{1}$$

Additionally, we compute the **Gender Gap (GG)**, which quantifies bias intensity by measuring the difference in resolution accuracy between male and female classifications:

$$GG = |RA_m - RA_f| \tag{2}$$

A higher **GG** value indicates stronger gender bias, as it reflects the disparity in classification performance between male and female subjects. By analyzing RA_{avg} and GG, we assess the impact of different debiasing strategies and determine which modality contributes more to gender bias in VLMs.

To determine whether the model classifies an image as male or female, we provide it with the image along with two prompts that differ only in the gender-related term (e.g., "The person in the image is male" vs. "The person in the image is female"). The model then generates logits (numerical representations) for both captions. We calculate the cosine similarity between the image's representation and each of the two captions. The caption with the higher similarity score determines the model's predicted gender for the image. For all the captions we used, refer to appendix. Here, we only analyse the results for the captions "The person in the image is male" vs. "The person in the image is female", as we found out that they show greater bias in the model.

4 Results

We obtain the results in table 1 using the prompts {'The person in the image is male', 'The person in the image is female'}. Our evaluation focuses on measuring gender bias in Vision-Language Models (VLMs) across different modalities. We analyze the impact of freezing text and vision modules on Resolution Accuracy (RA) and Gender Gap (GG).

4.1 Impact of Prompt Complexity

For results on all the prompts, refer appendix. We observe a significant difference in bias when using simple versus descriptive prompts:

- **Short prompts** (e.g., *"male/female," "he/she"*) maintain high RA$_m$, indicating that the model retains its ability to classify male images accurately.
- **Descriptive prompts** (e.g., *"The person in the image is male/female"*) lead to noticeable RA$_m$ drops, particularly in raw CLIP and when only partial finetuning is applied.

4.2 Effect of Freezing Text and Vision Encoders

We analyze the effect of freezing different model components across modalities:
(This analysis is only based on the prompts used for the results in Table 1)

- **Raw CLIP:** This setting shows us that the clip is inherently biased towards females beacuse of high GG and low RAm.
- **Freezing Vision Encoder:** RA$_m$ has improved drastically as compared to Raw CLIP model. This achieved the next least GG after None Frozen setting.
- **Freezing Text Encoder:** Even though there is improvement in RAm in this case, it is not as good as in the case where we freze the vision encoder.
- **None Frozen:** The best RA$_{avg}$ is achieved when both text and vision encoders are finetuned, minimizing GG and reducing overall bias.

4.3 Key Observations

Our analysis highlights the following patterns:

- **Text Encoders Exhibit More Bias:** When text is frozen, the gender gap increases, confirming that text embeddings contain stronger gender stereotypes than vision features.
- **Bias Exposure Through Descriptive Prompts:** Longer prompts introduce more gendered cues, making models rely on stereotypical associations.
- **Prompt Dependence:** Short prompts provide less biased information, whereas descriptive prompts expose deeper biases within the model.
- **Projection Layers:** On using CDA only on projection layers also shows a decrease in the bias, which shows us that the bias from the encoders is carried even to the projection layers.

Overall, these results suggest that targeted text adaptation is crucial for reducing gender bias in Vision-Language Models, while freezing text and debiasing vision only shows a minimal improvement. Refer to Figure 1 for visual intuition on how GG changes in different settings.

5 Conclusion

In this phase, we analyzed gender bias in CLIP by evaluating the contributions of text and vision modalities. Using Counterfactual Data Augmentation (CDA), we assessed the impact on mitigating bias. Our experiments reveal that in CLIP, the text encoder is the primary source of bias, as freezing text embeddings leads to a significant increase in the gender gap. We also found that simple prompts, such as "he/she," yield lower bias, whereas descriptive prompts amplify it.

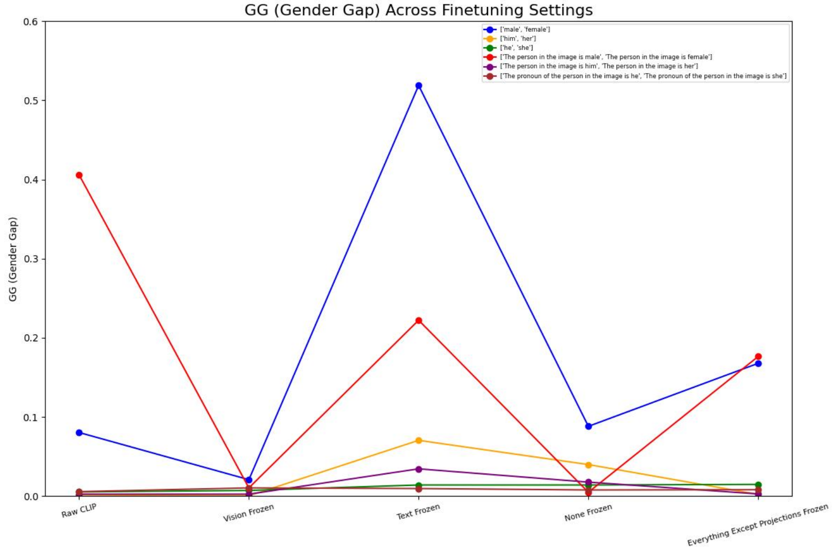

Figure 1: Visual intuition on how the Gender Gap (GG) changes across different debiasing settings.

For further phases, we plan to implement the same method on Pali-Gemma2 and also implement another method called Task Vector on both CLIP and Pali-Gemma2 and analyze whether these findings are generalizable by changing the model and the methods.

6 Appendix

Table 2: Results for Prompt: `['male', 'female']`

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Data: Anti-stereotypical Male				
No Finetuning	0.9930	0.9126	0.9528	0.0803
Vision Frozen	0.9883	0.9676	0.9779	0.0207
Text Frozen	1.0000	0.4811	0.7405	0.5189
None Frozen	0.9953	0.9072	0.9513	0.0881
Everything except Projections Frozen	0.9930	0.9000	0.9465	0.0930
Data: Anti-stereotypical Male and Female				
No Finetuning	0.9930	0.9126	0.9528	0.0803
Vision Frozen	0.9883	0.9829	0.9856	0.0054
Text Frozen	0.9930	0.9613	0.9771	0.0317
None Frozen	0.9883	0.9811	0.9847	0.0072
Everything except Projections Frozen	0.9930	0.9126	0.9528	0.0803

Table 3: Results for Prompt: `['him', 'her']`

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Data: Anti-stereotypical Male				
No Finetuning	0.9859	0.9874	0.9867	0.0015
Vision Frozen	0.9859	0.9856	0.9858	0.0003
Text Frozen	0.9930	0.9225	0.9577	0.0704
None Frozen	0.9930	0.9532	0.9731	0.0398
Everything except Projections Frozen	0.9859	0.9865	0.9862	0.0006
Data: Anti-stereotypical Male and Female				
No Finetuning	0.9859	0.9874	0.9867	0.0015
Vision Frozen	0.9859	0.9856	0.9858	0.0003
Text Frozen	0.9883	0.9775	0.9829	0.0108
None Frozen	0.9859	0.9847	0.9853	0.0012
Everything except Projections Frozen	0.9859	0.9874	0.9867	0.0015

Table 4: Results for Prompt: `['he', 'she']`

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Data: Anti-stereotypical Male				
No Finetuning	0.9836	0.9784	0.9810	0.0052
Vision Frozen	0.9812	0.9883	0.9848	0.0071
Text Frozen	0.9906	0.9766	0.9836	0.0140
None Frozen	0.9906	0.9766	0.9836	0.0140
Everything except Projections Frozen	0.9836	0.9775	0.9805	0.0061
Data: Anti-stereotypical Male and Female				
No Finetuning	0.9836	0.9784	0.9810	0.0052
Vision Frozen	0.9836	0.9892	0.9864	0.0056
Text Frozen	0.9836	0.9847	0.9841	0.0011
None Frozen	0.9883	0.9856	0.9869	0.0027
Everything except Projections Frozen	0.9859	0.9712	0.9785	0.0147

Table 5: Results for Prompt: `['The person in the image is male', 'The person in the image is female']`

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Data: Anti-stereotypical Male				
No Finetuning	0.5798	0.9856	0.7827	0.4058
Vision Frozen	0.9789	0.9892	0.9840	0.0103
Text Frozen	0.7535	0.9757	0.8646	0.2222
None Frozen	0.9859	0.9811	0.9835	0.0048
Everything except Projections Frozen	0.6479	0.9811	0.8145	0.3332
Data: Anti-stereotypical Male and Female				
No Finetuning	0.5798	0.9856	0.7827	0.4058
Vision Frozen	0.9859	0.9829	0.9844	0.0030
Text Frozen	0.8404	0.9964	0.9184	0.1560
None Frozen	0.9906	0.9766	0.9836	0.0140
Everything except Projections Frozen	0.7723	0.9486	0.8605	0.1763

Table 6: Results for Prompt: [’The person in the image is him’, ’The person in the image is her’]

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Data: Anti-stereotypical Male				
No Finetuning	0.9859	0.9883	0.9871	0.0024
Vision Frozen	0.9859	0.9883	0.9871	0.0024
Text Frozen	0.9930	0.9586	0.9758	0.0344
None Frozen	0.9906	0.9730	0.9818	0.0176
Everything except Projections Frozen	0.9836	0.9883	0.9859	0.0047
Data: Anti-stereotypical Male and Female				
No Finetuning	0.9859	0.9883	0.9871	0.0024
Vision Frozen	0.9859	0.9874	0.9867	0.0015
Text Frozen	0.9859	0.9775	0.9817	0.0084
None Frozen	0.9859	0.9865	0.9862	0.0006
Everything except Projections Frozen	0.9883	0.9856	0.9869	0.0027

Table 7: Results for Prompt: [’The pronoun of the person in the image is he’, ’The pronoun of the person in the image is she’]

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Data: Anti-stereotypical Male				
No Finetuning	0.9836	0.9892	0.9864	0.0056
Vision Frozen	0.9789	0.9892	0.9840	0.0103
Text Frozen	0.9906	0.9811	0.9858	0.0095
None Frozen	0.9906	0.9829	0.9867	0.0077
Everything except Projections Frozen	0.9789	0.9901	0.9845	0.0112
Data: Anti-stereotypical Male and Female				
No Finetuning	0.9836	0.9892	0.9864	0.0056
Vision Frozen	0.9859	0.9883	0.9871	0.0024
Text Frozen	0.9836	0.9856	0.9846	0.0020
None Frozen	0.9859	0.9865	0.9862	0.0006
Everything except Projections Frozen	0.9836	0.9784	0.9810	0.0052

Table 8: Results for Prompt: [’male’, ’female’]

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Data: Anti-stereotypical Male				
No Finetuning	0.9930	0.9126	0.9528	0.0803
Vision Frozen	0.9883	0.9676	0.9779	0.0207
Text Frozen	1.0000	0.4811	0.7405	0.5189
None Frozen	0.9953	0.9072	0.9513	0.0881
Everything except Projections Frozen	0.9930	0.9000	0.9465	0.0930
Data: Anti-stereotypical Male and Female				
No Finetuning	0.9930	0.9126	0.9528	0.0803
Vision Frozen	0.9883	0.9829	0.9856	0.0054
Text Frozen	0.9930	0.9613	0.9771	0.0317
None Frozen	0.9883	0.9811	0.9847	0.0072
Everything except Projections Frozen	0.9930	0.8252	0.9091	0.1677

Table 9: Results for Prompt: [’The person in the image is male’, ’The person in the image is female’]

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Data: Anti-stereotypical Male				
No Finetuning	0.5798	0.9856	0.7827	0.4058
Vision Frozen	0.9789	0.9892	0.9840	0.0103
Text Frozen	0.7535	0.9757	0.8646	0.2222
None Frozen	0.9859	0.9811	0.9835	0.0048
Everything except Projections Frozen	0.6479	0.9811	0.8145	0.3332
Data: Anti-stereotypical Male and Female				
No Finetuning	0.5798	0.9856	0.7827	0.4058
Vision Frozen	0.9859	0.9829	0.9844	0.0030
Text Frozen	0.8404	0.9964	0.9184	0.1560
None Frozen	0.9906	0.9766	0.9836	0.0140
Everything except Projections Frozen	0.7723	0.9486	0.8605	0.1763

Table 10: Results for Prompt: [’he’, ’she’]

Method	**RA**$_m$	**RA**$_f$	**RA**$_{avg}$	**GG**
Data: Anti-stereotypical Male				
No Finetuning	0.9836	0.9784	0.9810	0.0052
Vision Frozen	0.9812	0.9883	0.9848	0.0071
Text Frozen	0.9906	0.9766	0.9836	0.0140
None Frozen	0.9906	0.9766	0.9836	0.0140
Everything except Projections Frozen	0.9836	0.9775	0.9805	0.0061
Data: Anti-stereotypical Male and Female				
No Finetuning	0.9836	0.9784	0.9810	0.0052
Vision Frozen	0.9836	0.9892	0.9864	0.0056
Text Frozen	0.9836	0.9847	0.9841	0.0011
None Frozen	0.9883	0.9856	0.9869	0.0027
Everything except Projections Frozen	0.9859	0.9712	0.9785	0.0147

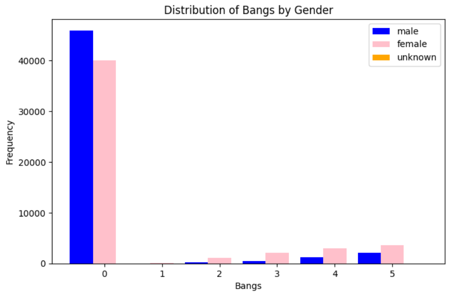

Figure 2: Distribution of Bangs by Gender.

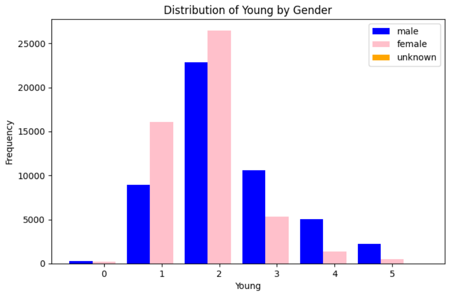

Figure 3: Distribution of Young by Gender.

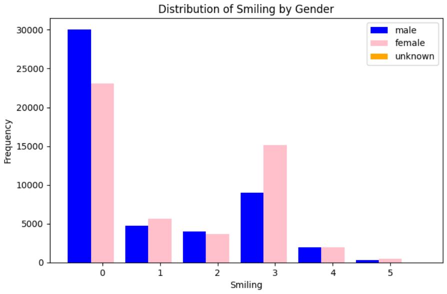

Figure 4: Distribution of Smiling by Gender.

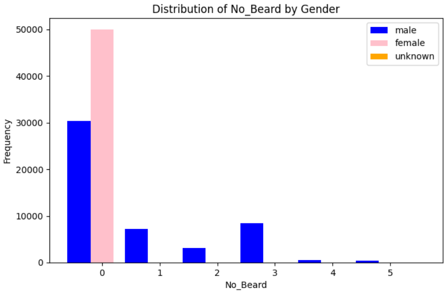

Figure 5: Distribution of Eye Glasses by Gender.

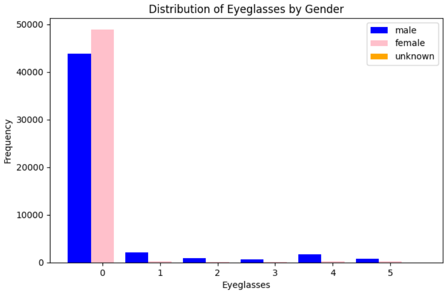

Figure 6: Distribution of No_Beard by Gender.

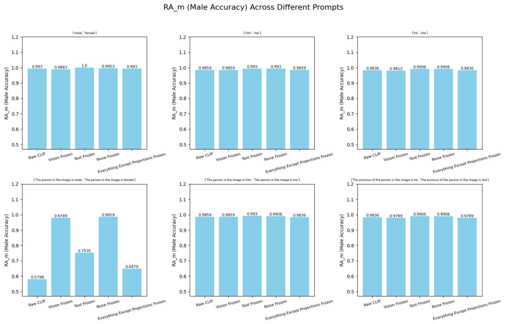

Figure 7: Resolution Accuracy - Male in different settings and prompts.

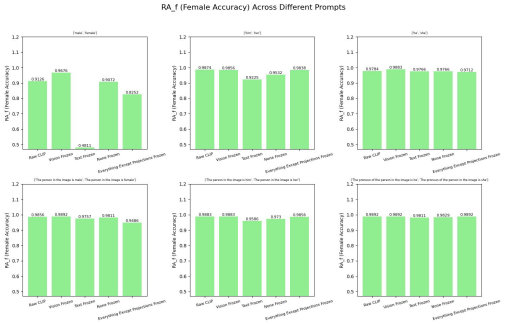

Figure 8: Resolution Accuracy - Female in different settings and prompts.

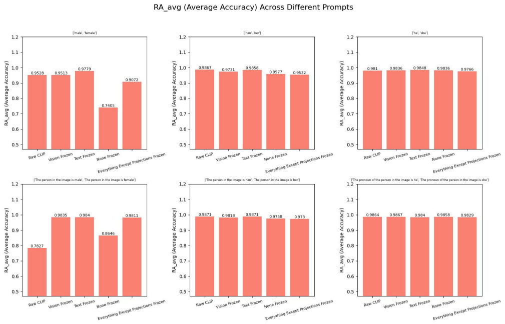

Figure 9: Resolution Accuracy - Average in different settings and prompts.