Freeze and Reveal: Exposing Modality Bias in Vision-Language Models

Kriti Madumadukala [*1] **Vysishtya Karanam** [*1] **Jahnavi Venkamsetty** [*1] **Vivek Hruday Kavuri** [*1]

Abstract

Vision-Language Models (VLMs) often inherit and amplify gender biases from their training data, yet the relative contributions of visual and textual modalities to this bias remain unclear. In this work, we systematically analyze the impact of text and vision on gender bias in VLMs by independently applying two debiasing techniques—Counterfactual Data Augmentation (CDA) and Task Vector—to each modality while freezing the other. Using CLIP and PaliGemma-2, we curate a gender-annotated CelebA-Dialogue dataset and evaluate bias mitigation effectiveness with the VisoGender benchmark. After evaluation on Visogender dataset, our findings reveal that, vision modality contributes towards more gender bias than the text modality and underscore the necessity of modality-specific interventions for equitable multimodal AI. This study provides a novel framework for disentangling multimodal bias and advancing fairness in VLMs. Further in appendix, we present our results on the other split of same data which we used for CDA and also from those results from which we could infer the dependence of prompt on revealing the gender biases.We release our code public at: https://github.com/vivekhruday05/VLM_Bias/

1. Introduction

The integration of visual and textual modalities in Vision-Language Models (VLMs) has led to remarkable advances in multimodal AI (Radford et al., 2021; Steiner et al., 2024), yet these models often inherit and even amplify gender biases present in their training data (Su et al., 2019). Such biases arise from stereotypical representations in both text and images, resulting in skewed perceptions that can propagate through downstream tasks. In this work, we address these challenges by applying targeted debiasing techniques—specifically Counterfactual Data Augmentation (CDA) (Wu & Dredze, 2020; Webster et al., 2021; Zmigrod et al., 2019) and Task Vector (Dige et al., 2024; Ilharco et al., 2023; Zhang et al., 2023) methods—to independently mitigate biases in the vision and text encoders. By curating a gender-annotated dataset and rigorously evaluating our methods using benchmarks like VisoGender (Hall et al., 2023), we aim to reveal which modality contributes more significantly to gender bias and to propose effective mitigation strategies. Ultimately, our study contributes to the development of fairer and more equitable multimodal AI systems.

2. Related work

Vision-Language Models (VLMs) such as CLIP and PaliGemma-2 have significantly advanced multimodal AI by integrating textual and visual modalities, enabling strong performance across diverse tasks. However, concerns have emerged regarding their tendency to inherit and amplify biases present in training data, particularly gender bias. This bias can stem from both text and image components, as language models trained on large-scale internet corpora frequently encode societal stereotypes, while image datasets may reinforce skewed gender representations by overrepresenting specific demographics in certain professions, emotions, or activities. The interaction between these modalities further complicates bias propagation, making it crucial to determine whether textual or visual elements contribute more significantly to gender bias in VLMs.

Several studies have attempted to quantify and mitigate bias in AI models. (Zhao et al., 2019) demonstrated how word embeddings reflect and reinforce societal biases, highlighting the problematic encoding of gender stereotypes in language representations. (Steed & Caliskan, 2021) analyzed multimodal bias in CLIP, revealing that gender and racial biases are amplified in the model's image-to-text mappings. (Gehman et al., 2020) introduced real-world benchmarks to measure societal biases in generative models, emphasizing the need for robust evaluation frameworks. (Moreira et al., 2024) explored debiasing techniques focused on text

[*]Equal contribution [1]IIIT Hyderabad, India. Correspondence to: Kriti Madumadukala <kriti.madumadukala@students.iiit.ac.in>, Vysishtya Karanam <vysishtya.karanam@students.iiit.ac.in>, Jahnavi Venkamsetty <venkata.venkamsetty@students.iiit.ac.in>, Vivek Hruday Kavuri <kavuri.hruday@research.iiit.ac.in>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

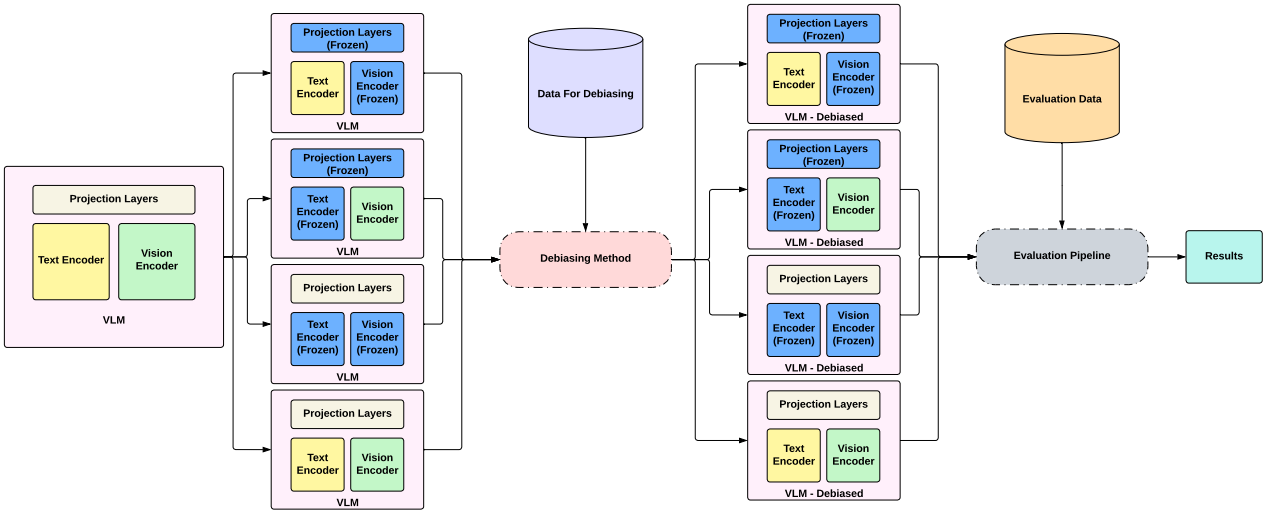

Figure 1. This figure shows the architecture of our method. First we start with a VLM, then in different settings, we freeze different modules and then we debias the unfrozen modules. Then we get debiased VLMs, where there may be partial debiasing (as in only a few modules) or full debiasing (as in not freezing any module and debiasing full model). Then we evaluate the models in these different settings and analyse the results.

prompts in multimodal models, indicating that interventions at the textual level can reduce bias to some extent but may not fully address the issue in vision-language interactions.

To mitigate gender bias, researchers have proposed several debiasing techniques, including Counterfactual Data Augmentation (CDA) and Task Vector methods. CDA works by synthetically generating counterfactual training data by swapping gendered terms (e.g., replacing ”he” with ”she”), thereby balancing gender representation in textual inputs (Zmigrod et al., 2020). While effective in NLP models, its application to VLMs remains underexplored. The Task Vector method, which leverages fine-tuned model parameter updates to create a debiasing transformation, has primarily been used for domain adaptation but has not been extensively studied in the context of bias mitigation in VLMs. Evaluating these techniques separately for text and vision components is essential for understanding their effectiveness in multimodal bias reduction.

Despite progress in bias mitigation, research gaps persist. Most studies analyze overall gender bias without isolating contributions from text and vision, leaving the primary source of bias unclear. Furthermore, existing benchmarks for gender bias evaluation in VLMs are limited, with few studies utilizing real-world datasets like VisoGender for rigorous assessment. Additionally, while CDA and Task Vector techniques have been explored in unimodal settings, their effectiveness in debiasing individual modalities in VLMs remains largely unexamined. Addressing these gaps is crucial for developing fairer and more unbiased multimodal AI systems.

This study aims to bridge these gaps by applying debiasing techniques to specific modalities and evaluating their impact. By independently implementing CDA and Task Vector methods on the text and vision components of VLMs, we aim to determine which modality contributes more to gender bias. Additionally, we will curate a gender-annotated CelebA-Dialogue dataset, allowing for a more precise analysis of gender bias in multimodal contexts. Unlike previous studies relying on pre-existing annotations, our approach ensures explicit annotation of gender and stereotypes. Rigorous evaluation will be conducted using VisoGender, a dedicated benchmark for assessing gender bias in VLMs. Furthermore, by comparing CLIP and PaliGemma-2, we will gain insights into how different VLM architectures propagate bias and respond to debiasing interventions.

By systematically evaluating debiasing strategies in a modality-specific manner, this research will contribute to ongoing efforts in fair AI development. The findings will offer critical insights into the origins of gender bias in VLMs and inform future research on effective bias mitigation techniques, ultimately guiding the development of more equitable multimodal AI models.

3. Dataset

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 ratings 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.

Figure 2. Example of raw dataset samples with annotations.

3.1. Data Pre-processing and Annotation

First, we require gender labels for every data point. To achieve this, we employ a rule-based automatic labeler. Specifically, we search for gender-related terms or pronouns such as *his/her*, *he/she*, *gentleman/lady*, and *male/female*. Based on the presence of these words, we classify the data point as either male or female. If none of these words appear, the annotator assigns the label *unknown*. This approach results in only 40 data points labeled as *unknown*, which is negligible compared to the dataset size, allowing us to prune them.

Next, we annotate the data points for stereotype classification. The dataset includes a rating from 0 to 5 for each data point across attributes {*Bangs, Smiling, No Beard, Young, Eye Glasses*}. Based on these ratings and predefined thresholds for stereotypical male and female characteristics, we label data points as either *stereotypical* or *anti-stereotypical*. These thresholds are determined by referring to prior publications and statistical insights from the dataset.

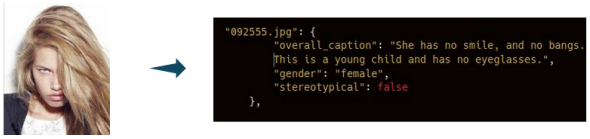

Figure 3. Data sample after Preprocessing

4. Methodology

Our main objective is to determine which modality—vision or text—contributes more to gender bias in our selected models. To achieve this, as shown in the Figure 1, we independently debias the encoder for each modality while keeping the rest of the model frozen, and then assess the overall bias using our evaluation metrics. The modality that, when debiased separately, leads to a greater reduction in bias is considered to be inherently more biased.

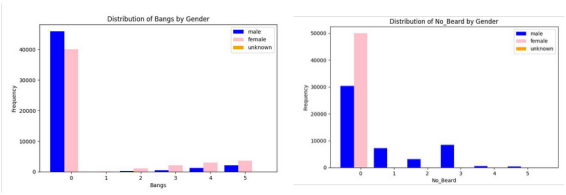

Figure 4. Distributions of different attributes across gender categories. The thresholds for stereotype classification are set based on the observed distributions of attributes such as *Bangs* and *No Beard*, ensuring alignment with statistical patterns in the dataset.

This approach allows us to isolate the bias contributions of each encoder and provides insights into which modality is a more significant source of bias in the integrated vision-language model. To achieve this, we use pre-existing debiasing methods that debias the whole model to independently debias the encoder for each modality while keeping the rest of the model frozen. The debiasing methods we plan to use are Counterfactual Data Augmentation (CDA) and Task Vector.

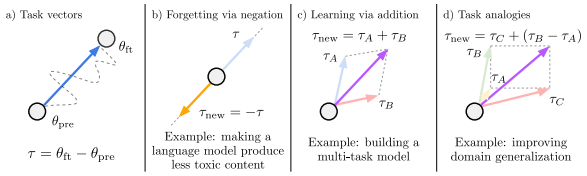

Figure 5. Applications of task vector, taken directly from (Ilharco et al., 2023)

4.1. Counter Factual Data Augmentation

As discussed in (Wu & Dredze, 2020; Webster et al., 2021; Zmigrod et al., 2019), Counterfactual Data Augmentation (CDA) is a technique that mitigates biases by incorporating counterfactual data into the training process. In this approach, the model is fine-tuned on augmented data that challenges stereotypical associations, which helps to attenuate biased representations.

We define counterfactual data as examples that contradict prevailing stereotypes. By augmenting these these anti-stereotypical examples, we hypothesize that the model will better recognize and handle non-stereotypical patterns, thus reducing inherent biases. Given that our methodology requires pre-existing debiasing mechanisms to independently address biases in the model's multimodal encoders, CDA is integrated as one of the experimental settings in our study.

4.2. Task Vector

As discussed in (Dige et al., 2024; Ilharco et al., 2023; Zhang et al., 2023), the Task Vector is derived by subtracting

the weights of a base model from those of a model fine-tuned on a specific task. This difference isolates the task-specific adjustments made during fine-tuning. Consequently, manipulating the Task Vector—for example, by negating it—can effectively "forget" the task-specific features, while adding it can reinforce them (see Figure 5).

In our approach, we first fine-tune the pre-trained base model using a dataset of stereotypical sentences, resulting in a biased model. We then compute the Task Vector by subtracting the base model's weights from those of the biased model. Finally, by applying the appropriately scaled negated Task Vector to the base model, we obtain a debiased model that mitigates the biases learned during fine-tuning.

5. Experiments

For CDA we use the anti-stereotypical examples from the dataset we annotated and fine-tune *openai/clip-vit-base-patch32*. We do this in 4 different settings, namely:

- **Vision Frozen:** In this setting, we freeze all the modules in a model except for the text encoder. There by only modifying the weights corresponding to the text encoder in the back propagation.
- **Text Frozen:** In this setting, we freeze all the modules in a model except for the vision encoder. There by only modifying the weights corresponding to the vision encoder in the back propagation.
- **None Frozen:** In this setting, we do not freeze any of the the modules in a model. There by modifying all the weights corresponding the model in the back propagation.
- **Only Projections Unfrozen:** In this setting, we freeze all the modules in a model except for the projection layers. There by only modifying the weights corresponding to the projection layers in the back propagation.

We use Nvidia Geforce 2080 Ti for finetuning the models on the anti-stereotypical data. We describe the evaluation pieline and the results in the upcoming sections.

6. Results

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:

Table 1. Results on the Visogender dataset in OO (Occupation Object) setting. High RA implies better performance and Low GG implies lesser Bias. First best performing metric is in **bold** next best performing is in *italic*

METHOD	RA_M	RA_F	RA_AVG	GG
RAW CLIP	0.9130	**0.9739**	0.9435	0.0609
VISION FROZEN	0.9130	*0.9652*	0.9391	0.0522
TEXT FROZEN	**0.9652**	*0.9652*	**0.9652**	**0.0000**
NONE FROZEN	**0.9652**	*0.9652*	**0.9652**	**0.0000**
ONLY PROJECTIONS UNFROZEN	*0.9304*	*0.9652*	*0.9478*	*0.0348*

Table 2. Results on the Visogender dataset in OP (Occupation Participant) setting. High RA implies better performance and Low GG implies lesser Bias. First best performing metric is in **bold** next best performing is in *italic*

METHOD	RA_M	RA_F	RA_AVG	GG
RAW CLIP	0.4087	**0.7130**	0.5609	0.3043
VISION FROZEN	0.4826	0.6522	0.5674	0.1696
TEXT FROZEN	*0.5391*	0.6217	*0.5804*	*0.0826*
NONE FROZEN	**0.5957**	*0.6609*	**0.6283**	**0.0652**
ONLY PROJECTIONS UNFROZEN	0.4261	0.6957	0.5609	0.2696

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

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.

We then obtain the model logits and it's preferences of gender for the data in Visogender Benchmark in different settings available, namely, Object-Occupation (OO) and Occupation-Participant (OP). After obtaining these gender preferences of the model and using the true labels of the dataset, we calculate the defined metrics and report in the further sub-sections. For more details about the prompt templates used for evaluation on Visogender, refer Appendix.

Tables 1 and 2 summarize the performance of our models on the VisoGender benchmark (Hall et al., 2023) under different debiasing settings for the Occupation-Object (OO)

and Occupation-Participant (OP) experiments, respectively. Our evaluation focuses on two key metrics: RA_avg, which represents the average recognition accuracy across genders, and GG, a measure of the gender gap (with lower values indicating reduced bias). We visualize our results in figures 6 and 7.

6.1. OO Experiments

In the OO setting (Table 1), the baseline Raw Clip model achieves an RA_avg of 0.9435 with a GG of 0.0609. Debiasing the text encoder (Vision Frozen) slightly lowers the average accuracy to 0.9391 while reducing the gap to 0.0522. In contrast, when the vision encoder is debiased (Text Frozen), the model not only achieves a higher RA_avg of 0.9652 but also completely eliminates the gender gap (GG = 0.0000). The None Frozen configuration produces identical results to Text Frozen, indicating that the inherent bias from the vision modality may be more effectively mitigated. The Only Projections Unfrozen setting yields intermediate results (RA_avg = 0.9478, GG = 0.0348), suggesting that simply training the projection layer without debiasing the encoders is less effective and this also shows that the bias from encoders is also carried to projection layers.

6.2. OP Experiments

For the OP setting (Table 2), the baseline Raw Clip model shows a lower RA_avg of 0.5609 accompanied by a much higher gender gap of 0.3043. Debiasing the text encoder (Vision Frozen) improves RA_avg modestly to 0.5674 and reduces GG to 0.1696, while debiasing the vision encoder (Text Frozen) further enhances performance (RA_avg = 0.5804, GG = 0.0826). Notably, when the entire model is left unfrozen (None Frozen), the model achieves the best performance with an RA_avg of 0.6283 and the lowest GG of 0.0652. The Only Projections Unfrozen setting, however, does not provide significant bias reduction (RA_avg = 0.5609, GG = 0.2696), but this slight reduction shows that the bias from encoders is also carried to projection layers.

6.3. Discussion

Our results indicate that independently debiasing the vision encoder has a more pronounced effect on reducing gender bias than debiasing the text encoder. In both the OO and OP experiments, configurations that involve debiasing (or leaving unfrozen) the vision modality lead to a substantial decrease in the gender gap. This suggests that the vision modality is a more significant source of bias within our integrated vision-language models. In particular, the complete elimination of the gender gap in the OO setting when the vision encoder is debiased underscores the effectiveness of this modality-specific approach.

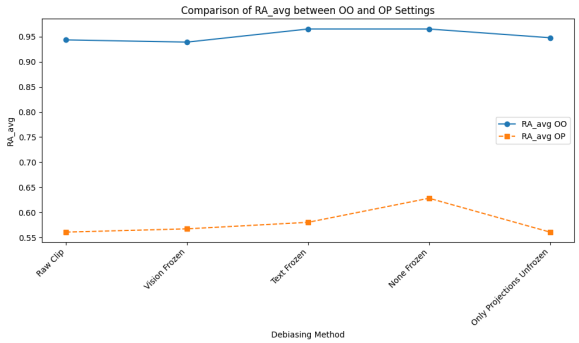

Figure 6. Plot of RA_{avg} in different settings.

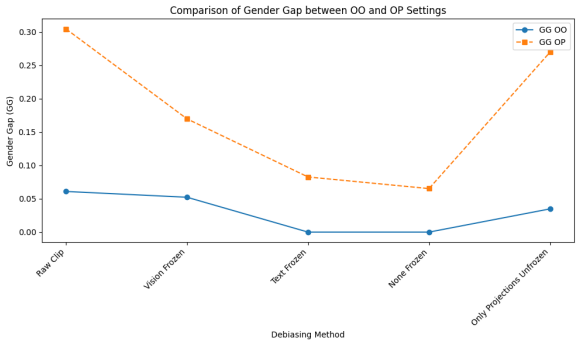

Figure 7. Plot of GG in different settings.

Overall, our findings highlight the importance of targeting the vision modality in debiasing strategies for vision-language models. By isolating the contributions of each encoder, we provide evidence that the vision component is more influential in propagating gender bias, and that appropriate debiasing of this modality can lead to a more balanced and fair model performance.

7. Conclusion

In this study, we introduced a framework to mitigate gender bias in vision-language models by isolating and independently debiasing the text and vision encoders. Using Counterfactual Data Augmentation (CDA) and Task Vector methods, we analyzed the contribution of each modality to overall bias. Our experimental results on the CelebA-Dialogue dataset and evaluations with the VisoGender benchmark reveal that the vision modality plays a more significant role in propagating gender bias. Configurations that debiased or left the vision encoder unfrozen led to a marked reduction or complete elimination of the gender gap, while debiasing the text encoder resulted in only modest improvements. These findings highlight the importance of targeted debiasing interventions in multimodal systems and provide actionable insights for designing fairer AI models. By demonstrat-

ing that modality-specific debiasing can effectively reduce bias, our work advances the understanding of gender bias in vision-language systems and lays the groundwork for future research. In future work, we will implement the Task Vector method on both models and evaluate its effectiveness using the VisoGender dataset, further assessing the potential of Task Vector-based interventions in reducing bias and enhancing model fairness.

References

Dige, O., Arneja, D., Yau, T. F., Zhang, Q., Bolandraftar, M., Zhu, X., and Khattak, F. K. Can machine unlearning reduce social bias in language models? In Dernoncourt, F., Preoţiuc-Pietro, D., and Shimorina, A. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 954–969, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.71. URL https://aclanthology.org/2024.emnlp-industry.71/.

Fitousi, D. Stereotypical processing of emotional faces: Perceptual and decisional components. *Frontiers in Psychology*, 12, 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.733432. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.733432.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models, 2020. URL https://arxiv.org/abs/2009.11462.

Hall, S. M., Abrantes, F. G., Zhu, H., Sodunke, G., Shtedritski, A., and Kirk, H. R. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution, 2023. URL https://arxiv.org/abs/2306.12424.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic, 2023. URL https://arxiv.org/abs/2212.04089.

Moreira, D. A. B., Ferreira, A. I., Silva, J., dos Santos, G. O., Pereira, L., Gondim, J. M., Bonil, G., Maia, H., da Silva, N., Hashiguti, S. T., dos Santos, J. A., Pedrini, H., and Avila, S. Fairpivara: Reducing and assessing biases in clip-based multimodal models, 2024. URL https://arxiv.org/abs/2409.19474.

Muthukumar, V., Pedapati, T., Ratha, N., Sattigeri, P., Wu, C.-W., Kingsbury, B., Kumar, A., Thomas, S., Mojsilovic, A., and Varshney, K. R. Understanding unequal gender classification accuracy from face images, 2018. URL https://arxiv.org/abs/1812.00099.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Steed, R. and Caliskan, A. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 701–713. ACM, March 2021. doi: 10.1145/3442188.3445932. URL http://dx.doi.org/10.1145/3442188.3445932.

Steiner, A., Pinto, A. S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., Qin, S., Ingle, R., Bugliarello, E., Kazemzadeh, S., Mesnard, T., Alabdulmohsin, I., Beyer, L., and Zhai, X. Paligemma 2: A family of versatile vlms for transfer, 2024. URL https://arxiv.org/abs/2412.03555.

Su, H., Shen, X., Zhang, R., Sun, F., Hu, P., Niu, C., and Zhou, J. Improving multi-turn dialogue modelling with utterance ReWriter. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 22–31, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1003. URL https://aclanthology.org/P19-1003/.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., and Petrov, S. Measuring and reducing gendered correlations in pre-trained models, 2021. URL https://arxiv.org/abs/2010.06032.

Wu, S. and Dredze, M. Are all languages created equal in multilingual BERT? In Gella, S., Welbl, J., Rei, M., Petroni, F., Lewis, P., Strubell, E., Seo, M., and Hajishirzi, H. (eds.), *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16. URL https://aclanthology.org/2020.repl4nlp-1.16/.

Zhang, J., Chen, S., Liu, J., and He, J. Composing parameter-efficient modules with arithmetic operations, 2023. URL https://arxiv.org/abs/2306.14870.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. Gender bias in contextualized word embeddings, 2019. URL https://arxiv.org/abs/1904.03310.

Zmigrod, R., Mielke, S. J., Wallach, H., and Cotterell, R. Counterfactual data augmentation for mitigating

gender stereotypes in languages with rich morphology. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1161. URL https://aclanthology.org/P19-1161/.

Zmigrod, R., Mielke, S. J., Wallach, H., and Cotterell, R. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology, 2020. URL https://arxiv.org/abs/1906.04571.

A. Appendix

A.1. VisoGender Prompt Templates for CLIP

For CLIP like models, we use the following sentence templates:

- For OP setting with Occupation first: The $OCCUPATION and $POSS_PRONOUN $PARTICIPANT
- For OP setting with Participant first: The $PARTICIPANT and $POSS_PRONOUN $OCCUPATION
- For OO setting: $NOM_PRONOUN is a $OCCUPATION

  Given these prompt templates, we fill in the Occupation and Participant with the actual gold labels and then give this filled prompt and the Image to the model and make the model predict the pronoun. We do the same by obtaining logits for the pronoun for male, female and neutral pronouns and taking the gender as the one which has highest magnitude of logits. Then, we calculate the evaluation metrics as stated in the results section. In OO setting, there is only one person and one object where as in OP setting there are 2 people. In OO section, it only predicts the gender of that one person present, but where as in OP section, it predicts the genders of both Participant and the main person in two different settings.

A.2. Evaluation of CLIP on same dataset's Validation Split

Before evaluating on the Visogender dataset, we evaluate our models on the Validation split of the dataset we have used to debias using CDA method. We present the same here.

Table 3. Results for Prompt: `['male', 'female']`

Method	$\mathbf{RA}_m$	$\mathbf{RA}_f$	$\mathbf{RA}_{avg}$	GG
Data: Anti-stereotypical Male				
Raw Clip	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
Only Projections Unfrozen	0.9930	0.9000	0.9465	0.0930
Data: Anti-stereotypical Male and Female				
Raw Clip	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
Only Projections Unfrozen	0.9930	0.9126	0.9528	0.0803

Publications supporting the threshold set for annotation of stereotypes https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.733432/full https://arxiv.org/abs/1812.00099.

Table 4. Results for Prompt: `['him', 'her']`

Method	$\mathbf{RA}_m$	$\mathbf{RA}_f$	$\mathbf{RA}_{avg}$	GG
Data: Anti-stereotypical Male				
Raw Clip	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
Only Projections Unfrozen	0.9859	0.9865	0.9862	0.0006
Data: Anti-stereotypical Male and Female				
Raw Clip	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
Only Projections Unfrozen	0.9859	0.9874	0.9867	0.0015

Table 5. Results for Prompt: `['he', 'she']`

Method	$\mathbf{RA}_m$	$\mathbf{RA}_f$	$\mathbf{RA}_{avg}$	GG
Data: Anti-stereotypical Male				
Raw Clip	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
Only Projections Unfrozen	0.9836	0.9775	0.9805	0.0061
Data: Anti-stereotypical Male and Female				
Raw Clip	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
Only Projections Unfrozen	0.9859	0.9712	0.9785	0.0147

Table 6. Results for Prompt: `['The person in the image is male', 'The person in the image is female']`

Method	$\mathbf{RA}_m$	$\mathbf{RA}_f$	$\mathbf{RA}_{avg}$	GG
Data: Anti-stereotypical Male				
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
Only Projections Unfrozen	0.6479	0.9811	0.8145	0.3332
Data: Anti-stereotypical Male and Female				
Raw Clip	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
Only Projections Unfrozen	0.7723	0.9486	0.8605	0.1763

Table 7. Results for Prompt: `['The person in the image is him', 'The person in the image is her']`

Method	$\mathbf{RA}_m$	$\mathbf{RA}_f$	$\mathbf{RA}_{avg}$	GG
Data: Anti-stereotypical Male				
Raw Clip	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
Only Projections Unfrozen	0.9836	0.9883	0.9859	0.0047
Data: Anti-stereotypical Male and Female				
Raw Clip	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
Only Projections Unfrozen	0.9883	0.9856	0.9869	0.0027

Table 8. Results for Prompt: `['The pronoun of the person in the image is he', 'The pronoun of the person in the image is she']`

Method	$\mathbf{RA}_m$	$\mathbf{RA}_f$	$\mathbf{RA}_{avg}$	GG
Data: Anti-stereotypical Male				
Raw Clip	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
Only Projections Unfrozen	0.9789	0.9901	0.9845	0.0112
Data: Anti-stereotypical Male and Female				
Raw Clip	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
Only Projections Unfrozen	0.9836	0.9784	0.9810	0.0052

Table 9. Results for Prompt: ['male', 'female']

Method	$\mathbf{RA}_m$	$\mathbf{RA}_f$	$\mathbf{RA}_{avg}$	GG
Data: Anti-stereotypical Male				
Raw Clip	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
Only Projections Unfrozen	0.9930	0.9000	0.9465	0.0930
Data: Anti-stereotypical Male and Female				
Raw Clip	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
Only Projections Unfrozen	0.9930	0.8252	0.9091	0.1677

Table 10. 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				
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
Only Projections Unfrozen	0.6479	0.9811	0.8145	0.3332
Data: Anti-stereotypical Male and Female				
Raw Clip	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
Only Projections Unfrozen	0.7723	0.9486	0.8605	0.1763

Table 11. Results for Prompt: ['he', 'she']

Method	RA_m	RA_f	RA_{avg}	GG
Data: Anti-stereotypical Male				
Raw Clip	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
Only Projections Unfrozen	0.9836	0.9775	0.9805	0.0061
Data: Anti-stereotypical Male and Female				
Raw Clip	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
Only Projections Unfrozen	0.9859	0.9712	0.9785	0.0147

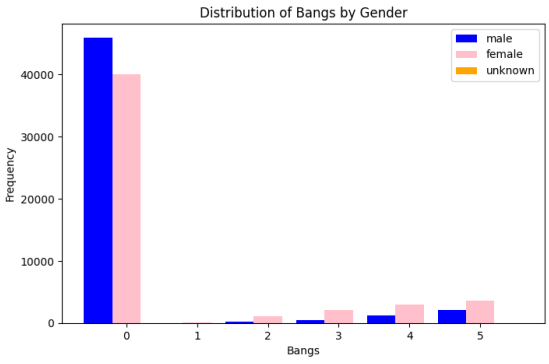

Figure 8. Distribution of Bangs by Gender.

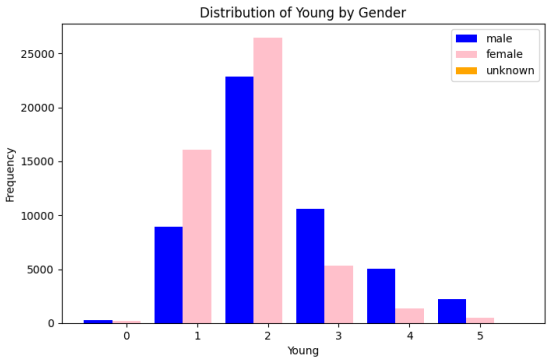

Figure 9. Distribution of Young by Gender.

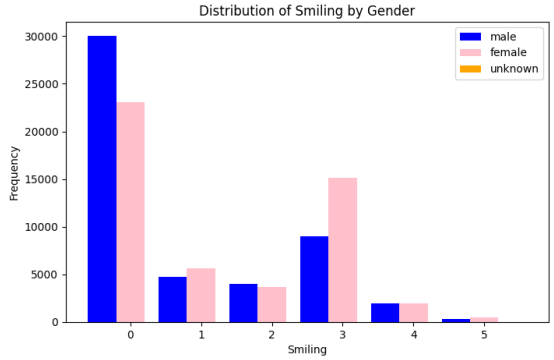

Figure 10. Distribution of Smiling by Gender.

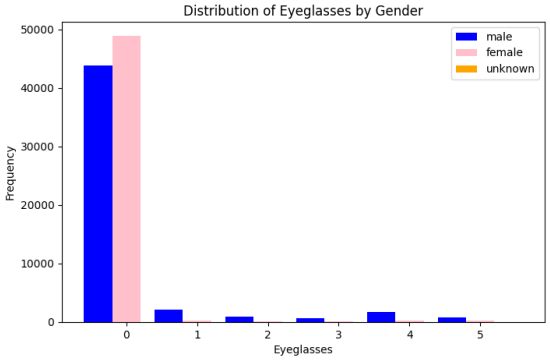

Figure 11. Distribution of Eye Glasses by Gender.

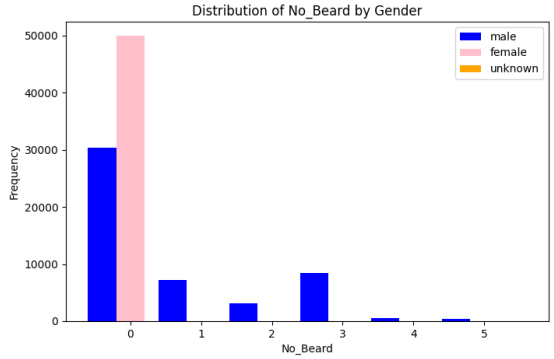

Figure 12. Distribution of No_Beard by Gender.

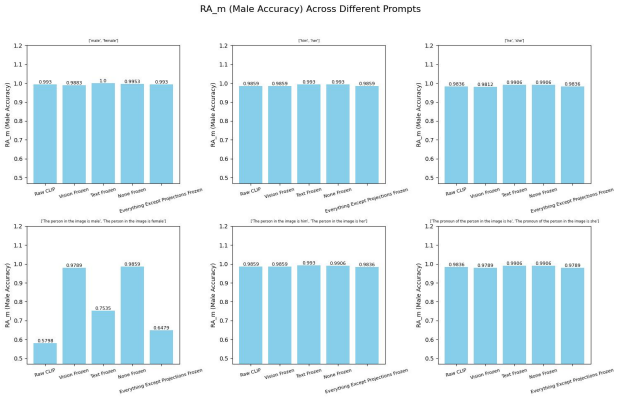

Figure 13. Resolution Accuracy - Male in different settings and prompts.

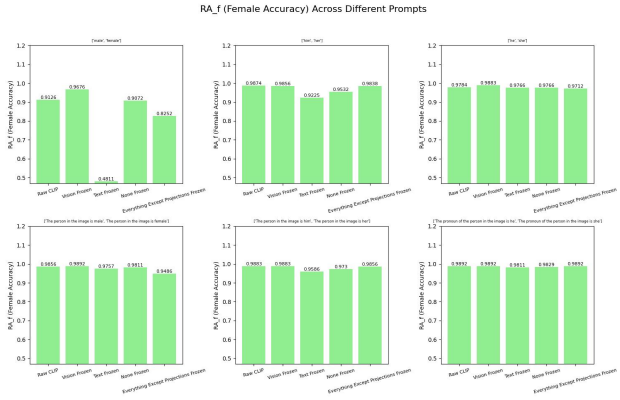

Figure 14. Resolution Accuracy - Female in different settings and prompts.

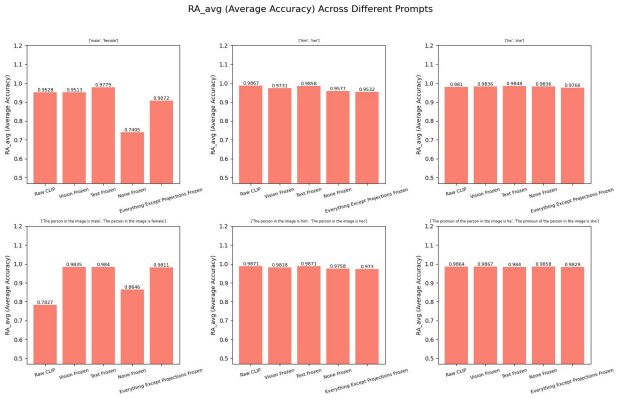

Figure 15. Resolution Accuracy - Average in different settings and prompts.