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) have achieved impressive performance across multimodal tasks, yet they often inherit and amplify gender biases from their training data. In this work, we investigate the relative contributions of the vision and text encoders to such biases and apply targeted debiasing strategies—Counterfactual Data Augmentation (CDA) and Task Vector methods—to mitigate them independently. Inspired by data-efficient approaches in hate speech classification, we introduce a novel metric, *Degree of Stereotypicality* (DoS), and a corresponding debiasing method, *Data Augmentation Using DoS* (DAUDoS), to reduce bias with minimal computational cost. We curate a gender-annotated dataset and evaluate all methods on the VisoGender benchmark to quantify improvements and identify the dominant source of bias. Our results show that both CDA and DAUDoS yield measurable gains in fairness, with nuanced differences in modality-specific contributions. We release our code publicly 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), 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.

Then, Inspired from Nejadgholi et al. (2022) and Garg et al. (2025), we also propose a metric called Degree of Stereotypicality (DoS) and an efficient Debiasing technique Data Augmentation Using DoS (DAUDoS). Using this method too, we independently mitigate biases in the vision and text encoders and evaluate the same on Visogender benchmark to reveal which modality contributes more significantly to gender bias.

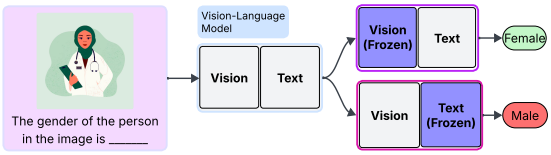

Figure 2. Different modalities posses different level of biases. We aim to show which modality exhibits more.

Our contributions are summarized as follows:

- We address gender bias in Vision-Language Models (VLMs) by applying targeted debiasing strategies—specifically Counterfactual Data Augmentation (CDA) and Task Vector arithmetic—to independently intervene on the vision and text encoders.
- We curate a gender-annotated dataset and evaluate our debiasing methods using the VisoGender benchmark, enabling a rigorous comparison of the relative contributions of visual and textual modalities to gender bias.
- Inspired by prior work on data-efficient generalization in hate speech detection, we propose a novel metric

[*]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.

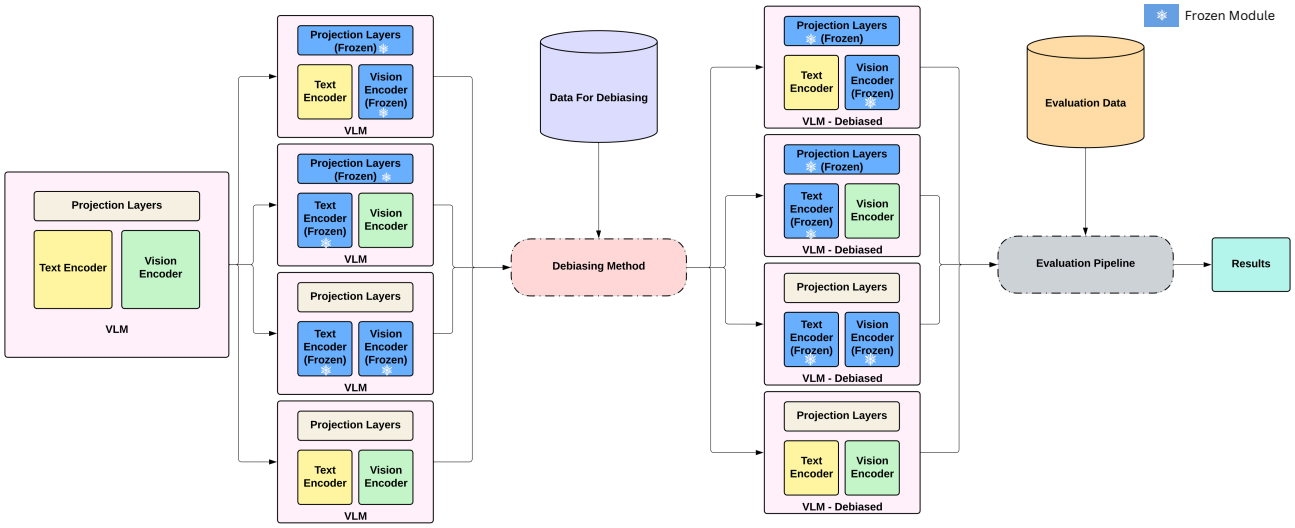

Figure 1. This figure shows the architecture of our method. First we start with a VLM, then in different settings, we freeze different modules and then we debias the unfrozen modules. Then we get debiased VLMs, where there may be partial debiasing (as in only a few modules) or full debiasing (as in not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

called *Degree of Stereotypicality* (DoS), which quantifies the stereotypical nature of multimodal samples.

- We introduce an efficient debiasing technique, *Data Augmentation Using DoS* (DAUDoS), which selectively augments the training set based on DoS scores to reduce stereotypical correlations in both modalities with minimal computational overhead.

- We demonstrate that both CDA and DAUDoS, when applied independently to vision and text encoders, lead to measurable reductions in gender bias. Our analysis reveals which modality plays a more dominant role in propagating harmful stereotypes within VLMs.

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 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.

Further, training on all counterfactual examples can be computationally expensive and time-consuming. To address this, prior works such as Nejadgholi et al. (2022) and Garg et al. (2025) propose approaches for improving generalization in

hate speech classification while relying on fewer annotated examples. These methods leverage Concept Activation Vectors (CAVs) and introduce a novel metric, the *Degree of Explicitness*, which quantifies the explicit nature of hateful content. By assigning explicitness scores to samples, they selectively fine-tune models on a curated subset of training instances, thereby enhancing efficiency without compromising performance.

Despite such advancements in natural language processing (NLP) for hate speech detection, their extension to the domain of multi-modal AI—especially for bias mitigation—remains relatively underexplored. Inspired by these works, we propose a novel metric termed the *Degree of Stereotypicality* (DoS), which quantifies how strongly a sample exhibits stereotypical associations. Building on this, we introduce a data-efficient bias mitigation strategy called *Data Augmentation using DoS* (DAUDoS), which enables targeted augmentation based on stereotypicality scores, thereby reducing computational overhead while maintaining or improving model fairness and robustness.

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 3. 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.

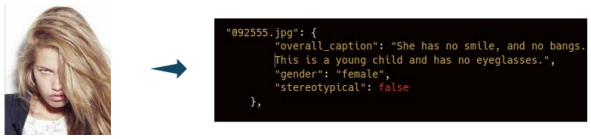

Figure 4. Data sample after Preprocessing

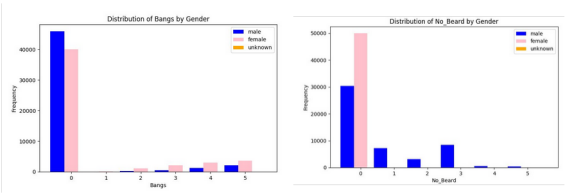

Figure 5. Distributions of different attributes across gender categories. The thresholds for stereotype classification are set based on the observed distributions of attributes such as *Bangs* and *No Beard*, ensuring alignment with statistical patterns in the dataset.

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.

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 Weighted Task Vector.

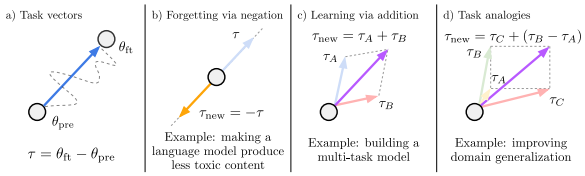

Figure 6. Applications of task vector, taken directly from (Ilharco et al., 2023)

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. To enhance flexibility in debiasing strength, we introduce a *weighted Task Vector method*, controlled by two hyperparameters: α and `blend`. Specifically, we adjust the original weights using:

$$W_{\text{debiased}} = W_{\text{original}} - ((1 - \texttt{blend}) \cdot \alpha) \cdot \Delta W_{\text{task}} \quad (1)$$

Here, α controls the overall intensity of debiasing, while $\texttt{blend} \in [0, 1]$ interpolates between the original and fully debiased model. A higher `blend` retains more of the original model's behavior, while a lower value emphasizes debiasing more strongly.

To identify optimal hyperparameters, we perform a random search over $\alpha \in [0.1, 1.0]$ and $\texttt{blend} \in [0.0, 1.0]$, guided by a loss that balances accuracy and fairness:

$$\mathcal{L} = -\text{RA}_{\text{avg}} + \lambda_{\text{gap}} \cdot \text{GenderGap} \quad (2)$$

where RA_{avg} is the average resolution accuracy across male and female identities, and $\text{GenderGap} = |\text{RA}_m - \text{RA}_f|$ penalizes disparity. This formulation promotes both high performance and equitable behavior by controlling for bias introduced during fine-tuning.

4.3. Data Augmentation Using DoS (DAUDoS)

In this section, we introduce *Data Augmentation Using DoS (DAUDoS)*, a targeted debiasing strategy that leverages the stereotypicality of samples to perform efficient fine-tuning. The overall process is illustrated in Figure 7.

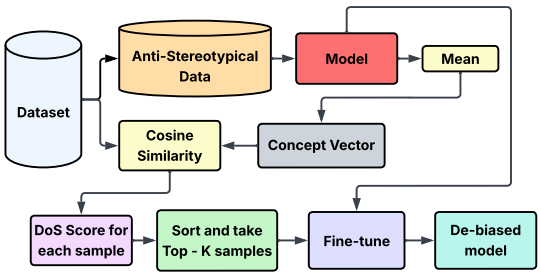

Figure 7. Diagram depicting the method Data Augmentation Using DoS (DAUDoS)

The key idea behind DAUDoS is to assign a *Degree of Stereotypicality* (DoS) score to each sample in the dataset. To do so, we begin by constructing a small set of anti-stereotypical samples. These are fed into a pre-trained model to obtain embeddings, from which we compute a *Concept Activation Vector (CAV)*. Formally, if $\{\mathbf{z}_i\}_{i=1}^n$ are the model embeddings of the anti-stereotypical samples, the concept vector $\mathbf{v}_{\text{CAV}}$ is computed as their mean:

$$\mathbf{v}_{\text{CAV}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i.$$

Next, for each input sample x, we obtain its model embedding $\mathbf{z}_x$ and compute its cosine similarity with $\mathbf{v}_{\text{CAV}}$:

$$\text{DoS}(x) = \cos(\mathbf{z}_x, \mathbf{v}_{\text{CAV}}).$$

This DoS score captures how closely the sample aligns with the concept of anti-stereotypicality: higher scores indicate lower stereotypicality, and vice versa.

Once scores are assigned, we sort all training samples by their DoS values and select the top-K most stereotypical samples for fine-tuning. This allows us to focus training on the subset of data that contributes most to bias, thereby making the process compute-efficient. These selected samples are used to fine-tune the model, leading to a debiased version as shown in Figure 7.

By guiding the data augmentation process with DoS, DAUDoS minimizes training cost while retaining effectiveness in bias mitigation across modalities.

5. Experiments

For CDA we use the anti-stereotypical examples from the dataset we annotated and fine-tune *openai/clip-vit-base-patch32*. Then for task vector, we used the stereotypical data to finetune the model and obtain task vector. In DAUDoS, we selceted the samples based on the scores irrespective of what the label of the sample is (whether it's stereotypical

or anti-stereotypical). We do these methods as discussed previously, 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 the 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:

$$RA_{avg} = \frac{RA_m + RA_f}{2} \tag{3}$$

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

$$GG = |RA_m - RA_f| \tag{4}$$

A higher GG indicates stronger gender bias, whereas a lower GG suggests more balanced performance across genders.

Our evaluation considers model logits and their corresponding gender preferences on the Visogender benchmark (Hall et al., 2023) in two settings: **Occupation-Object (OO)** and **Occupation-Participant (OP)**. In the **OO** setting, each instance involves a single individual paired with an occupational cue; the model is tasked with assigning the correct gender label based solely on the visual representation and the occupational context. Conversely, the **OP** setting presents a more complex scenario in which each sample includes two individuals with different roles, requiring the model to simultaneously predict the gender of multiple participants. This dual framework enables us to assess the model's ability to handle both isolated and relational gender cues, thereby providing a comprehensive view of its fairness in gender classification.

After obtaining the gender preference scores and using the true labels of the dataset, we compute RA_{avg} and GG for various debiasing configurations. In the following subsections, we report the results for the CLIP and Paligemma2 models.

6.1. CLIP Results

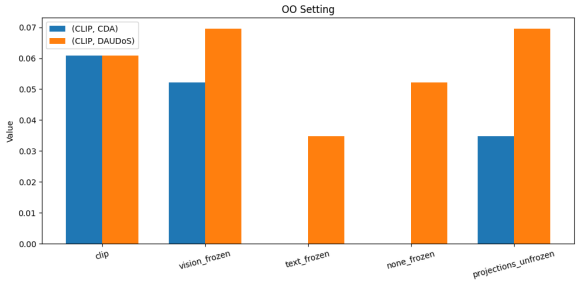

Figure 8. Plot of GG in different debiasing configurations for CLIP. Blue denotes CDA and Orange Denotes DAUDoS

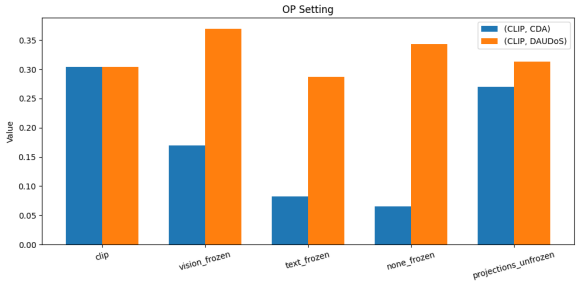

Figure 9. Plot of GG in different debiasing configurations for CLIP. Blue denotes CDA and Orange Denotes DAUDoS

Table 1 summarizes the performance of CLIP under different debiasing configurations. In the OO experiments, the *Raw Clip* baseline achieves an RA_{avg} of 0.9435 and a moderate GG of 0.0609. Freezing the text encoder alone (Vision Frozen) slightly reduces RA_{avg} to 0.9391 and decreases GG to 0.0522. Notably, when the vision encoder is debiased (Text Frozen), CLIP achieves an RA_{avg} of 0.9652 with the gender gap completely eliminated ($GG = 0.0000$). A configuration where both encoders are left trainable (None Frozen) mirrors these outcomes. The Projections Not Frozen setting yields intermediate results, with

$RA_{avg} = 0.9478$ and $GG = 0.0348$.

In the OP experiments (right columns of Table 1), the Raw CLIP model demonstrates a much lower RA_{avg} of 0.5609 and a high GG of 0.3043. Debiasing the text encoder (Vision Frozen) improves RA_{avg} to 0.5674 and reduces GG to 0.1696. Further improvement occurs when the vision encoder is debiased (Text Frozen), yielding $RA_{avg} = 0.5804$ and $GG = 0.0826$. Finally, allowing both encoders to update (None Frozen) provides the highest RA_{avg} (0.6283) with the lowest observed GG (0.0652).

Figure 9 displays a subplot of GG across the different debiasing configurations for CLIP, clearly illustrating that interventions aimed at debiasing the vision encoder (text_frozen setting) are particularly effective in lowering the gender gap. Hence, the more biased encoder in CLIP is vision encoder.

6.2. Paligemma2 Results

Table 2 shows the performance of the Paligemma2 model under similar conditions. In the CDA experiments, configurations such as Vision Frozen and None Frozen achieve very high RA_{avg} (approximately 0.9739–0.9870) while maintaining a very low gender gap (e.g., $GG = 0.0087$ for Vision Frozen). For the DAUDoS setting, while the RA_{avg} remains high (around 0.9435–0.9600), we need to note that we have only used 1/3rd of the samples we used for CDA.

Figure 11 provides a subplot of GG for the Paligemma2 model, reinforcing the trend that debiasing the text modality (vision_frozen) is particularly effective in reducing gender bias. Hence the more biased modality in PaliGemma2 is the text modality.

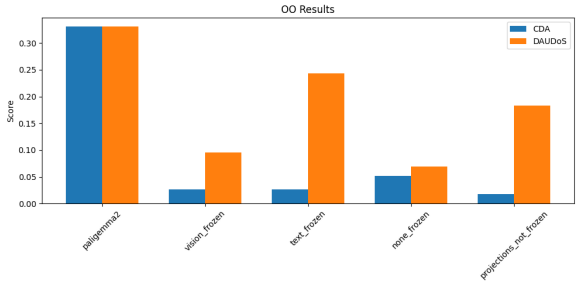

Figure 10. Plot of GG in different debiasing configurations for Paligemma2. Blue denotes CDA and Orange Denotes DAUDoS

6.3. Discussion

Our experiments reveal that independently debiasing encoders of different modalities has a more pronounced effect on reducing gender bias than debiasing model in whole. In both OO and OP settings, for CLIP configurations that involve targeting the vision encoder by allowing it to learn from debiased signals—consistently yield lower gender gap

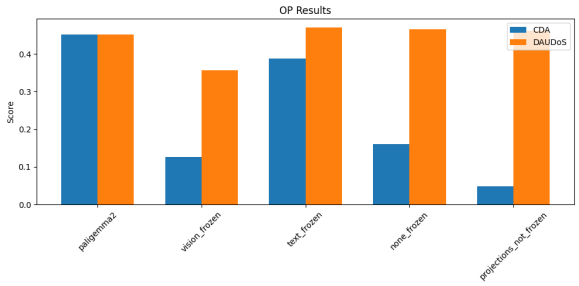

Figure 11. Plot of GG in different debiasing configurations for Paligemma2. Blue denotes CDA and Orange Denotes DAUDoS

values with minimal impact on average resolution accuracy, whereas in PaliGemma2, the text encoder shows similar results. This may be due to the fact that the Vision encoder of PaliGemma2 only has 0.5B parameters whereas the text encoder has 2.5B parameters. Hence, When we freeze the text encoder, we are not debiasing most of the parameters. But, in clip, where we have almost equal number of parameters for both modalities, debiasing only vision has a more pronounced effect, hence leading to the conclusion that vision encoder is more biased.

These findings underscore the importance of focusing on independent modalities when developing debiasing strategies for vision-language models. In contrast, intervening at the projection layer only provides a limited bias reduction, reflecting the fact that biases in the encoder are propagated downstream. Overall, our results suggest that modality-specific debiasing is critical for achieving a more balanced and fair model performance.

For further details on the prompt templates and evaluation procedures applied on the Visogender benchmark, please refer to Appendix A.

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), 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 in CLIP and text modality in PaliGemma2 plays a more significant propagating gender bias. Then, the method we introduce, DAUDoS, also show the same trend, supporting generalization across methods. These findings highlight the importance of targeted debiasing interventions in multimodal systems and provide actionable insights for designing fairer AI models. By demonstrating that modality-specific debiasing can effec-

Table 1. CLIP results in OO vs. OP settings. High RA implies better performance, low GG implies less bias.

	CDA							
Freeze Type	**RA_m**	**RA_f**	**RA_avg**	**GG**	**RA_m**	**RA_f**	**RA_avg**	**GG**
	OO				**OP**			
Raw Clip	0.9130	0.9739	0.9435	0.0609	0.4087	0.6522	0.5609	0.3043
Vision Frozen	0.9130	**0.9652**	0.9391	0.0522	0.4826	0.6522	0.5674	0.1696
Text Frozen	**0.9652**	**0.9652**	**0.9652**	**0.0000**	*0.5391*	0.6217	*0.5804*	*0.0826*
None Frozen	**0.9652**	**0.9652**	**0.9652**	**0.0000**	**0.5957**	*0.6609*	**0.6283**	**0.0652**
Projections	*0.9304*	**0.9652**	*0.9478*	*0.0348*	0.4261	**0.6957**	0.5609	0.2696
Task Vector (α = 0.5611, blend = 0.7812)								
Vision Frozen	*0.1739*	**0.7478**	**0.4609**	0.5739	0.1000	*0.0217*	0.0609	**0.0783**
Text Frozen	**0.6261**	0.2348	*0.4304*	*0.3913*	**0.5565**	**0.2217**	*0.3891*	0.3348
None Frozen	0.0696	*0.2609*	0.1652	**0.1913**	*0.2957*	0.0087	0.1522	*0.2870*
Projections Not Frozen	**0.6261**	0.2348	*0.4304*	*0.3913*	**0.5565**	**0.2217**	**0.3981**	0.3348
DAUDoS								
Vision Frozen	0.9130	**0.9826**	*0.9478*	0.0696	0.3826	*0.7522*	0.5674	0.3696
Text Frozen	**0.9391**	*0.9739*	**0.9565**	**0.0348**	**0.4565**	0.7435	*0.6000*	**0.2870**
None Frozen	*0.9304*	**0.9826**	**0.9565**	*0.0522*	*0.4391*	**0.7826**	**0.6109**	0.3435
Projections Not Frozen	0.9043	*0.9739*	0.9391	0.0696	0.4000	0.7130	0.5565	*0.3130*

Table 2. Paligemma2 results in OO vs. OP settings. High RA implies better performance, low GG implies less bias.

	CDA							
Freeze Type	**RA_f**	**RA_m**	**RA_avg**	**GG**	**RA_f**	**RA_m**	**RA_avg**	**GG**
	OO				**OP**			
Raw Paligemma	0.7913	0.4609	0.6261	0.3304	0.9043	0.4522	0.6783	0.4522
Vision Frozen	**0.9913**	**0.9826**	**0.9870**	**0.0087**	0.7174	0.7826	0.7500	*0.0652*
Text Frozen	0.4174	0.3913	0.4043	*0.0261*	0.6522	0.4739	0.5630	0.1783
None Frozen	0.9783	0.9696	0.9739	**0.0087**	**0.7609**	**0.8609**	**0.8109**	0.1000
Projections Not Frozen	*0.9870*	*0.9783*	*0.9826*	**0.0087**	*0.7391*	*0.7957*	*0.7674*	**0.0565**
DAUDoS								
Vision Frozen	*0.9000*	**0.9870**	*0.9435*	*0.0870*	**0.6478**	*0.8739*	**0.7609**	*0.2261*
Text Frozen	0.4783	0.6696	0.5739	0.1913	0.5000	0.8043	0.6522	0.3043
None Frozen	**0.9261**	**0.9870**	**0.9565**	**0.0609**	0.5217	**0.9087**	*0.7152*	0.3870
Projections Not Frozen	0.5304	*0.6783*	0.6043	0.1478	*0.5391*	0.7391	0.6391	**0.2000**

tively 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.

Garg, S., Kavuri, V. H., Shroff, G., and Mishra, R. Ktcr: Improving implicit hate detection with knowledge transfer driven concept refinement, 2025. URL https://arxiv.org/abs/2410.15314.

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.

Jung, H., Jang, T., and Wang, X. A unified debiasing approach for vision-language models across modalities and tasks. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 21034–21058. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/254404d551f6ce17bb7407b4d6b3c87b-Paper-Conference.pdf.

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.

Nejadgholi, I., Fraser, K., and Kiritchenko, S. Improving generalizability in implicitly abusive language detection with concept activation vectors. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5517–5529, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.378. URL https://aclanthology.org/2022.acl-long.378/.

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. Selective Feature Imputation for Debiasing (SFID)

To mitigate bias in vision-language models (VLMs) such as CLIP, we apply the Selective Feature Imputation for Debiasing (SFID) method Jung et al. (2024), a lightweight and post-hoc debiasing strategy that avoids retraining.

Given frozen representations of a modality (either vision or text), SFID identifies a small subset of features that are strongly predictive of a sensitive attribute (e.g., gender). This is achieved by training a simple classifier—such as a random forest—on the embeddings to predict the attribute. The classifier's feature importances are then used to identify the top-k biased dimensions.

SFID performs a *low-confidence imputation* on these biased dimensions by overwriting their values with the mean values computed from the least confident predictions of the classifier. Intuitively, these low-confidence samples represent "neutral" points in the embedding space, helping preserve semantics while mitigating bias.

Formally, let $E \in \mathbb{R}^{n \times d}$ denote the matrix of d-dimensional embeddings for n samples, and let $I \subset [1, d]$ be the indices of the top-k biased features. SFID computes imputation values μ_i for each $i \in I$ from the subset of samples where the classifier confidence is below a threshold τ. The debiased embedding matrix E' is defined as:

$$E'_{j,i} = \begin{cases} \mu_i & \text{if } i \in I \\ E_{j,i} & \text{otherwise} \end{cases}$$

We apply SFID separately to each modality (image or text) by extracting CLIP embeddings and measuring downstream attribute prediction accuracy before and after imputation. A reduction in attribute classification accuracy and disparity indicates successful bias mitigation.

In our experiments, we evaluate gender bias by computing resolution accuracy across subgroups (e.g., male vs female) and tracking the gender gap. We demonstrate that SFID reduces the gender gap across various CLIP configurations, including scenarios where the vision encoder, text encoder, or projection layers are selectively frozen.

A.3. Evaluation of CLIP on same dataset's Validation Split for CDA

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.

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

Method	RA_m	RA_f	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

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

Method	RA_m	RA_f	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	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

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	$\mathbf{RA}_m$	$\mathbf{RA}_f$	$\mathbf{RA}_{avg}$	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 11. 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

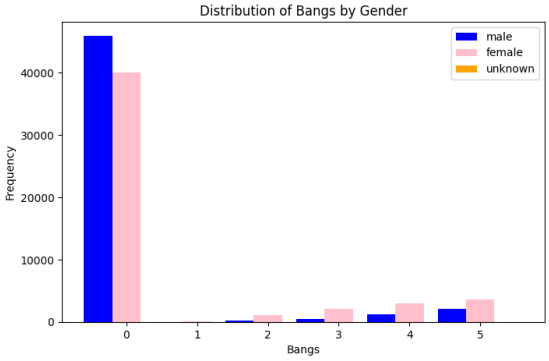

Figure 12. Distribution of Bangs by Gender.

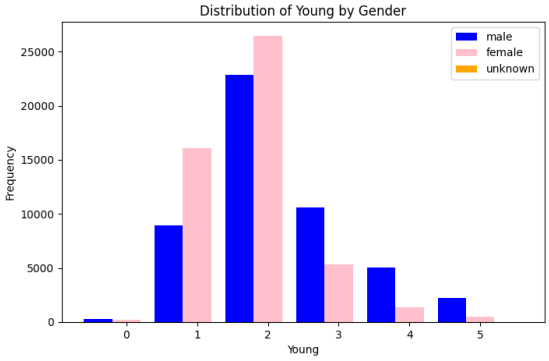

Figure 13. Distribution of Young by Gender.

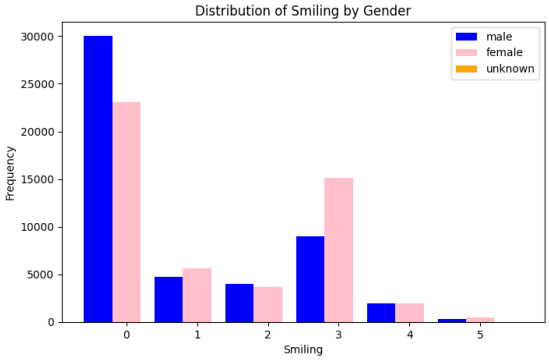

Figure 14. Distribution of Smiling by Gender.

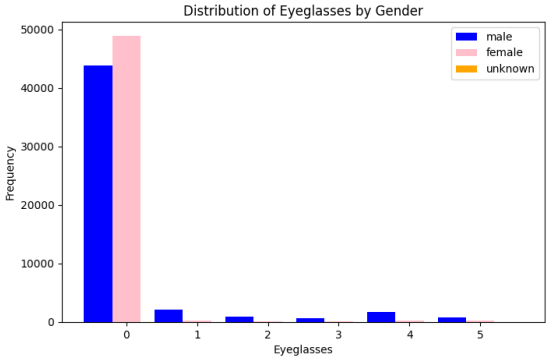

Figure 15. Distribution of Eye Glasses by Gender.

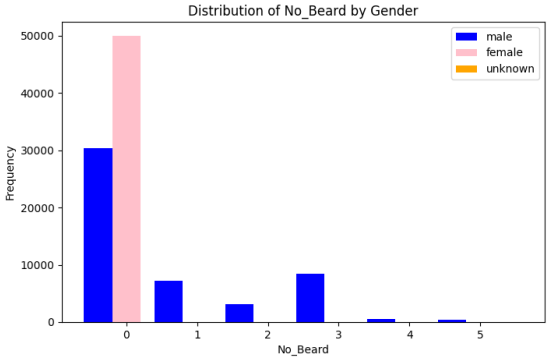

Figure 16. Distribution of No_Beard by Gender.

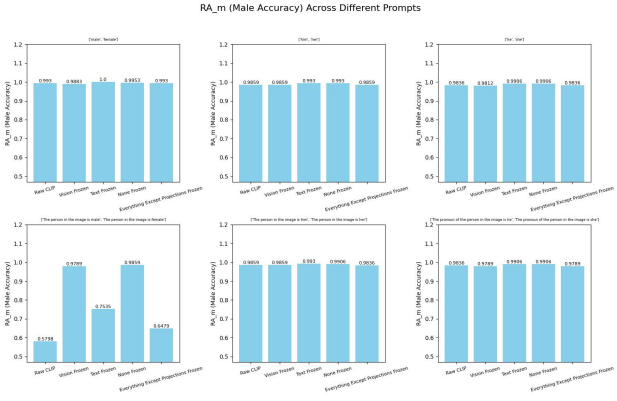

Figure 17. Resolution Accuracy - Male in different settings and prompts.

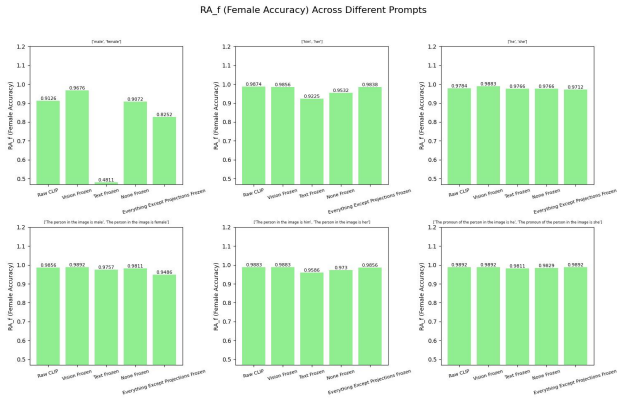

Figure 18. Resolution Accuracy - Female in different settings and prompts.

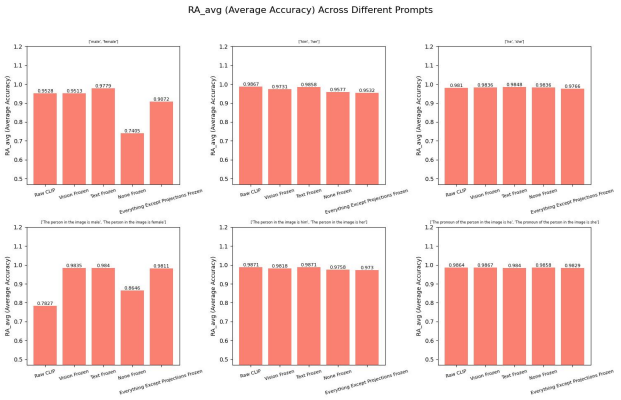

Figure 19. Resolution Accuracy - Average in different settings and prompts.