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

Vivek Hruday Kavuri, Vysishtya Karanam, Venkata Jahnavi Venkamsetty,
Kriti Madumadukala, Lakshmipathi Balaji Darur, Ponnurangam Kumaraguru
IIIT Hyderabad
{kavuri.hruday, lakshmipathi.balaji}@research.iiit.ac.in,
{vysishtya.karanam, venkata.venkamsetty, kriti.madumadukala}
@students.iiit.ac.in, pk.guru@iiit.ac.in

Abstract

Vision-Language Models (VLMs) achieve impressive multimodal performance but often inherit gender biases from their training data. This bias might be coming from both the vision and text modalities. In this work, we dissect the contributions of vision and text backbones to these biases by applying targeted debiasing—Counterfactual Data Augmentation (CDA) and Task Vector methods. 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 CDA reduces the gender gap by 6% and DAUDoS by 3% but using only one-third the data. Both methods also improve the model's ability to correctly identify gender in images by 3%, with DAUDoS achieving this improvement using only almost one-third of training data. From our experiments, we observed that CLIP's vision encoder is more biased whereas PaliGemma2's text encoder is more biased. By identifying whether the bias stems more from the vision or text encoders, our work enables more targeted and effective bias mitigation strategies in future multi-modal systems. We release our code public at https://github.com/vivekhruday05/VLM_bias

1 Introduction

The integration of visual and textual modalities in VLMs has led to remarkable advances in multimodal AI (Radford et al., 2021; Steiner et al., 2024; Li et al., 2022, 2023; Achiam et al., 2023; Team et al., 2023). VLMs have demonstrated exceptional capabilities across various tasks, including image retrieval (Xue et al., 2022; Bai et al., 2023), captioning (Li et al., 2022, 2023; Liu et al., 2024; Steiner et al., 2024). However these models often inherit gender biases present in their training data (Su et al., 2019) thus making them not suitable/reliable for real world deployment. Such biases also arise from stereotypical representations in both text and images, resulting in skewed perceptions that can propagate through downstream tasks.

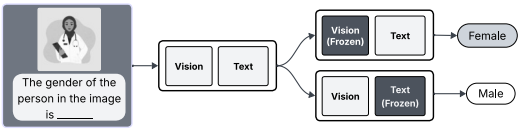

Figure 1: Different modalities posses different level of bias. We aim to show which one exhibits more bias.

In this work, we address these challenges by applying targeted debiasing techniques for both modalities. Specifically for a given VLM we debias a particular modality sub-module on a curated dataset and evaluate it for gender bias using VisoGender (Hall et al., 2023) to determine impact of each modality on gender bias. For this purpose we use the CelebA-Dialog dataset (Jiang et al., 2021) and curate the samples from the same. We annotate the data for gender based on the pronouns used in the caption and stereotypicality based on the statistical distribution of the data and insights from previous works (Fitousi, 2021; Muthukumar et al., 2018). To determine if a particular modality has higher influence in model's bias we evaluate it across multiple methods on our dataset. (i) We use CDA (Wu and Dredze, 2020; Webster et al., 2021; Zmigrod et al., 2019) a technique that mitigates bias by incorporating counterfactual data into the training process. (ii) We adapt Task Vector Unlearning (Dige et al., 2024; Ilharco et al., 2023;

Proceedings of the First Interdisciplinary Workshop on OMMM associated with RANLP 2025,
pages 17–26, Varna, Bulgaria, Sep 8-10, 2025.
https://doi.org/10.26615/978-954-452-101-1-002

Zhang et al., 2023) for debiasing. (iii) We propose a data-efficient debiasing approach, DAUDoS. We propose and do this for both CLIP-like similarity score based models and captioning type models and evaluate them across different methods. We consistently observe across multiple methods that CLIP's vision encoder is more biased compared to text encoder and in case of PaliGemma2, it's text encoder is more biased when compared to vision encoder.

In summary our key contributions are as follows:

- We propose a modality-targeted debiasing framework that applies CDA and Task-Vector methods separately to vision and text encoders to pinpoint each modality's bias.
- We curate a gender-annotated dataset for this analysis and evaluate our debiasing methods using the VisoGender benchmark.
- We propose DoS and introduce DAUDoS, lightweight debiasing methods that reduce gender bias on VisoGender with minimal overhead.

2 Related work

Bias in VLMs. 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 biases (Abdollahi et al., 2024; Darur et al., 2024; Xiao et al., 2024; Wolfe et al., 2023) 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 over representing specific demographics in certain professions, emotions, or activities. The interaction between these modalities further complicates the propagation of bias, making it crucial to determine whether textual or visual elements contribute more significantly to gender bias in VLMs. Previous works such as (Weng et al., 2024) focus on causal mediation to trace and mitigate gender bias in GLIP, showing image features contribute most and proposing input-level blurring to reduce bias. There are also works such as (Srinivasan and Bisk, 2022) which deal with bias measurement to multimodal models, revealing compounded intra and cross-modal stereotypes in VL-BERT. In contrast to these, our work targets a particular modality to find out which of the modalities contribute to a greater gender bias and whether they differ across different models and methods.

Bias Evaluation. Several studies have attempted to quantify and mitigate bias in AI models. Prior work has shown that word embeddings encode and perpetuate gender stereotypes in language representations (Zhao et al., 2019), that multimodal models like CLIP amplify both gender and racial biases in their image-to-text mappings (Steed and Caliskan, 2021). There are also existing real-world benchmarks which measure societal biases in generative models, emphasizing the need for robust evaluation frameworks (Gehman et al., 2020). 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 (Moreira et al., 2024).

Debiasing Techniques. To mitigate gender bias, researchers have proposed several debiasing techniques, including 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) and Task Vector (Ilharco et al., 2023) is an unlearning method which has it's roots originated from unlearing literature but also used in bias mitigation (Dige et al., 2024). While effective in NLP models, its application to VLMs remains underexplored.

Data-Efficient Debiasing. Training on all counterfactual examples can be computationally expensive and time-consuming. To address this, prior works (Nejadgholi et al., 2022; 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. Inspired by these advances in NLP, we extend these ideas to the multi-modal setting and 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 DuaDOS, which enables targeted augmentation based on stereotypicality scores. This approach reduces computational overhead while maintaining or improving model fairness and robustness in multimodal AI systems.

3 Dataset

We use the CelebA-Dialog dataset (Jiang et al., 2021) 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. Since, we require gender for each of the data point, both for applying our methods and evaluation, we annotate the gender and describe the process in the following subsections. We also need whether a data-point is stereotypical or anti-stereotypical, so that we can use for CDA. Hence, we also annotate that attribute and describe the process in the following subsections. An example of how initial data looks like is shown in Table 1.

Table 1: Examples of raw dataset samples with annotations. Each image is associated with both attribute-wise and overall captions, along with a numeric rating vector indicating the prominence of each attribute (e.g., bangs, eyeglasses, beard, smile, age) in order.

Image	
Bangs	He has no bangs at all. *Rating: 0*
Eyeglasses	There are no eyeglasses on the face. *Rating: 0*
Beard	This gentleman doesn't have any beard at all. *Rating: 0*
Smiling	This gentleman looks serious with no smile on his face. *Rating: 0*
Age	This person looks very old. *Rating: 5*
Overall Caption	This man in his eighties has no mustache, no fringe, and no smile. He is not wearing any eyeglasses.

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 male or female. If none of these words appears, the annotator assigns the label *unknown*. This approach results in only 40 data points labeled as *unknown*, which is negligible compared to the size of the dataset, 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 (Fitousi, 2021; Muthukumar et al., 2018). An example of a data point after the annotation is shown in Table 2.

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 2, 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

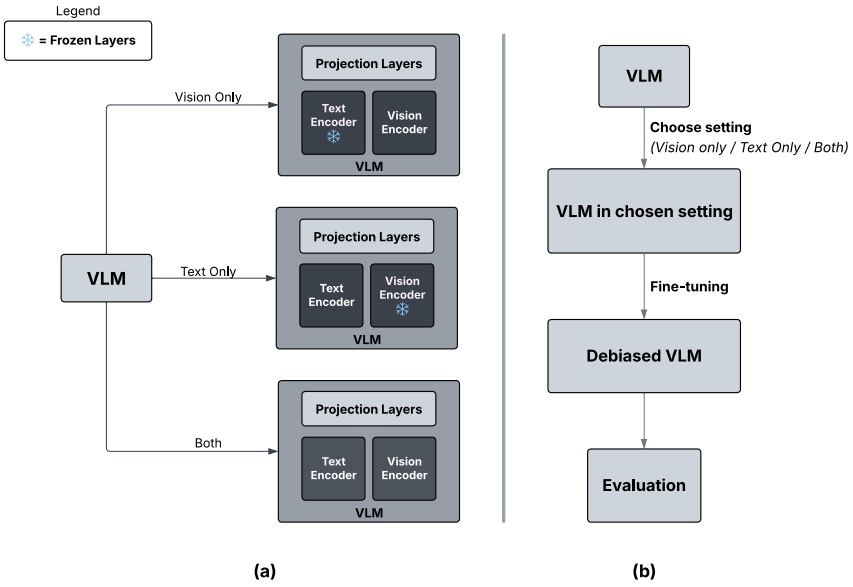

Figure 2: (a) Shows different layers that will be frozen in different settings we experiment in. (b) Shows an overall pipeline of our architecture. "Choose setting" means choosing a setting from one of the settings shown in (a).

Table 2: Data sample after preprocessing. Gender and stereotype labels are added based on rule-based and attribute rating analysis, respectively. Remaining attributes such as the ratings and individual captions are discarded.

Image	
Gender	Female
Stereotypical	False
Overall Caption	She has no smile and no bangs.This is a young child who has no eyeglasses.

into which modality is a more significant source of bias in the integrated VLM. 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 use are CDA and Weighted Task Vector.

4.1 Counter Factual Data Augmentation

As discussed in (Wu and 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 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 $\texttt{blend}$ 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:

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

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 DAUDoS, a targeted debiasing strategy that leverages the stereotypicality of samples to perform efficient fine-tuning. The overall process is illustrated in Figure 3.

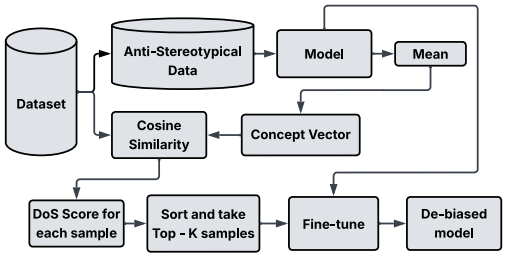

Figure 3: Depicting the method Data Augmentation Using DoS (DAUDoS). We first compute a concept vector from anti-stereotypical samples. Then, each dataset sample is scored based on its similarity to this vector, giving its Degree of Stereotypicality (DoS). The most stereotypical samples (more similarity with concept vector or score nearer to 1) are selected for fine-tuning, allowing targeted debiasing with minimal data.

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:

$$\mathbf{v}_{\text{CAV}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i. \tag{3}$$

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

$$\text{DoS}(x) = \cos(\mathbf{z}_x, \mathbf{v}_{\text{CAV}}). \tag{4}$$

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 3.

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 selected the samples based on the scores irrespective of what the label of the sample is (whether it is stereotypical or anti-stereotypical). We do these methods as discussed previously, in 4 different settings, namely:

Text only. In this setting, we freeze all the modules in a model except for the text encoder and projection layers related to text modality. There by only modifying the weights corresponding to the text encoder in the back propagation.

Vision Only. In this setting, we freeze all the modules in a model except for the vision encoder and projection layers related to vision modality. There by only modifying the weights corresponding to the vision encoder in the back propagation.

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

6 Results

To quantify gender bias in VLMs, as proposed in (Darur et al., 2024), 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:

$$RA_{avg} = \frac{RA_m + RA_f}{2} \tag{5}$$

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

$$GG = |RA_m - RA_f| \tag{6}$$

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

Table 3 summarizes the performance of CLIP under different debiasing configurations. In the OO experiments, the *Raw Clip* baseline achieves an RA_{avg} of 0.94 and a moderate GG of 0.06. Debiasing the text encoder alone (text only) has almost same RA_{avg} 0.94 and decreases GG to 0.052. Notably, when the vision encoder is debiased (vision only), CLIP achieves an RA_{avg} of 0.96 with the gender gap completely eliminated ($GG = 0.0000$). A configuration where both encoders are left trainable (both) mirrors the outcome same as that of the case when the vision modality is debiased.

In the OP experiments (right columns of Table 3), the Raw CLIP model demonstrates a much lower accuracy compared to OO setting with RA_{avg} 0.56 and a high GG of 0.30. Debiasing the text encoder (text only) improves RA_{avg} to 0.57 and reduces GG to 0.17. Further improvement occurs when the vision encoder is debiased (vision only), yielding $RA_{avg} = 0.58$ and $GG = 0.08$. Finally, allowing both encoders to update (both) provides the highest RA_{avg} (0.63) with the lowest observed GG (0.06).

Figure 4 and Figure 5 display a plot of GG across the different debiasing configurations for

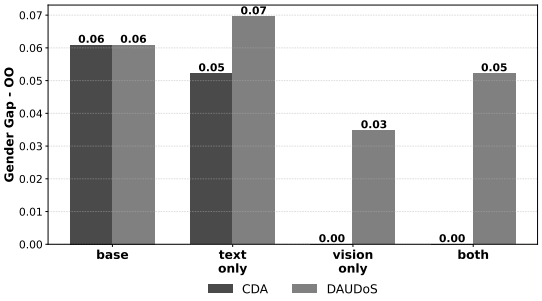

Figure 4: GG scores for **OO** setting in CLIP across debiasing configurations. Vision debiasing yields the least bias (GG = 0.0 by CDA, 0.03 by DAUDoS), similar to full model debiasing (GG = 0.0 by CDA, 0.05 by DAUDoS), indicating greater bias in the vision modality.

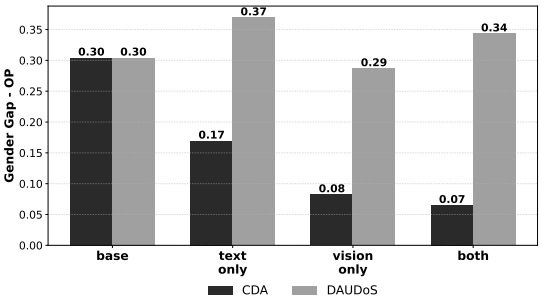

Figure 5: GG scores for **OP** setting in CLIP across debiasing configurations. Vision debiasing shows lowest bias (GG = 0.08 by CDA, 0.27 by DAUDoS), close to full model debiasing (GG = 0.07 by CDA, 0.34 by DAUDoS), again suggesting higher bias in the vision modality.

Table 3: Modality-targeted debiasing in CLIP under OO and OP settings. High RA implies better performance, low GG implies less bias. Debiasing the vision encoder in CLIP (Vision Only) achieves the highest RA_{avg} (0.97) with $GG = 0.00$, indicating vision contributes most bias.

	CDA							
Freeze Type	RA_m	RA_f	RA_{avg}	**GG**	RA_m	RA_f	RA_{avg}	**GG**
	OO				**OP**			
Raw Clip	0.91	0.97	0.94	0.06	0.41	0.65	0.56	0.30
Text Only	0.91	**0.97**	0.94	0.05	0.48	0.65	0.57	0.17
Vision Only	**0.97**	**0.97**	**0.97**	**0.00**	*0.54*	0.62	*0.58*	*0.08*
Both	**0.97**	**0.97**	**0.97**	**0.00**	**0.60**	*0.66*	**0.63**	**0.07**
Task Vector (α = 0.56, blend = 0.78)								
Text Only	*0.17*	**0.75**	**0.46**	0.57	0.10	*0.02*	0.06	**0.08**
Vision Only	**0.63**	0.23	*0.43*	*0.39*	**0.56**	**0.22**	*0.39*	0.33
Both	0.07	*0.26*	0.17	**0.19**	*0.30*	0.01	0.15	*0.29*
DAUDoS								
Text Only	0.91	**0.98**	*0.95*	0.07	0.38	*0.75*	0.57	0.37
Vision Only	**0.94**	*0.97*	**0.96**	**0.03**	**0.46**	0.74	*0.60*	**0.29**
Both	*0.93*	**0.98**	**0.96**	*0.05*	*0.44*	**0.78**	**0.61**	0.34

Table 4: Modality-targeted debiasing in PaliGemma2 under OO and OP settings. High RA implies better performance, low GG implies less bias. Debiasing the text encoder in PaliGemma2 (text only) yields $RA_{\text{avg}} = 0.99$ with $GG = 0.01$, showing text is the primary bias source.

	CDA							
Freeze Type	RA_f	RA_m	RA_{avg}	**GG**	RA_f	RA_m	RA_{avg}	**GG**
	OO				**OP**			
Raw Paligemma	0.79	0.46	0.63	0.33	0.90	0.45	0.68	0.45
Text Only	**0.99**	**0.98**	**0.99**	**0.01**	0.72	0.78	0.75	*0.07*
Vision Only	0.42	0.39	0.40	*0.03*	0.65	0.47	0.56	0.18
Both	0.98	0.97	0.97	**0.01**	**0.76**	**0.86**	**0.81**	0.10
DAUDoS								
Text Only	*0.90*	**0.99**	*0.94*	*0.09*	**0.65**	*0.87*	**0.76**	*0.23*
Vision Only	0.48	0.67	0.57	0.19	0.50	0.80	0.65	0.30
Both	**0.93**	**0.99**	**0.96**	**0.06**	0.52	**0.91**	*0.72*	0.39

CLIP, clearly illustrating that interventions aimed at debiasing the vision encoder (vision only setting) are particularly effective in lowering the gender gap. Hence, the more biased encoder in CLIP is vision encoder. We can observe this result consistently across methods.

6.2 Paligemma2 Results

Table 4 shows the performance of the Paligemma2 model under similar conditions. In the CDA experiments, configurations such as "text only" and "both" achieve very high RA_{avg} (approximately 0.97–0.99) while maintaining a very low gender gap (e.g., GG=0.01 for text only). For the DAUDoS setting, while the RA_{avg} remains high (around 0.94–0.96), it is important to note that these results were obtained using only one-third of the dataset. This aligns with our objective of achieving competitive performance using minimal data—demonstrating that selective sampling is both efficient and effective. Using the entire dataset would defeat the purpose of our sorting and data reduction strategy.

In the OP experiments (right columns of Table 4), the Raw model demonstrates similar accuracy compared to OO setting with RA_{avg} 0.68 and a high GG of 0.45. Debiasing the text encoder (text only) improves RA_{avg} to 0.75 and reduces GG to 0.07. But, notably no further improvement occurs when the vision encoder is debiased (vision only), yielding $RA_{avg} = 0.56$ and $GG = 0.18$. Finally, allowing both encoders to update (both) provides the highest RA_{avg} (0.81) but the Gender Gap GG of (0.06) is still higher than the gender gap observed in case of text only setting.

Figure 6 and Figure 7 provide a plot of GG for

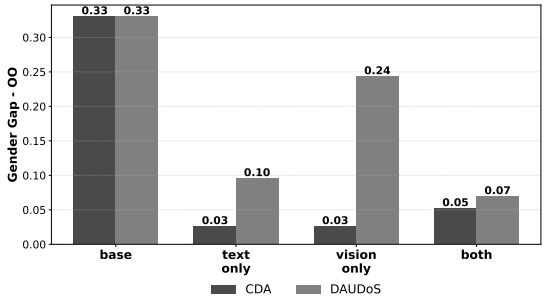

Figure 6: GG scores for **OO** setting across debiasing configurations for Paligemma2. Text debiasing yields lowest bias (GG = 0.03 by CDA, 0.10 by DAUDoS), similar to full model debiasing (GG = 0.05 by CDA, 0.07 by DAUDoS), suggesting higher bias in text modality.

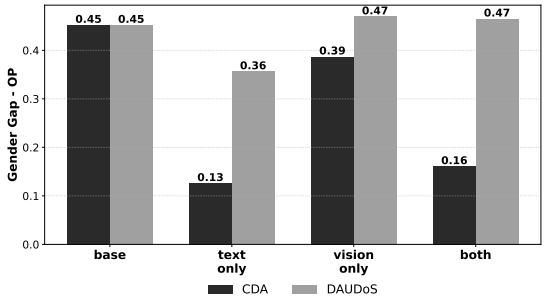

Figure 7: GG scores for **OP** setting across debiasing configurations for Paligemma2. Text debiasing gives lowest bias (GG = 0.13 by CDA, 0.36 by DAUDoS), close to full model debiasing (GG = 0.16 by CDA, 0.47 by DAUDoS), again pointing to text as the more biased modality.

the Paligemma2 model, reinforcing the trend that debiasing the text modality (Text Only) is particularly effective in reducing gender bias. Hence the more biased modality in PaliGemma2 is the text modality. We can observe this result consistently across methods.

7 Discussion

Our study investigates gender bias in VLMs by independently debiasing the text and vision encoders using methods like CDA and Task Vectors. Experiments on the CelebA-Dialogue dataset and evaluations with the VisoGender benchmark reveal that targeting individual modalities is more effective than intervening at the model level. In CLIP, debiasing the vision encoder yields lower gender gaps with minimal impact on accuracy—likely due to the balanced parameter sizes across modalities. In contrast, PaliGemma2's larger text encoder (2.5B parameters vs. 0.5B for vision) makes debiasing the text modality more impactful.

The findings also underscore that modality-specific debiasing leads to better bias mitigation than strategies applied post-encoder, such as projection layer adjustments, which only offer limited improvements. Our proposed DAUDoS method further supports this trend, demonstrating the generalizability of our approach across models and settings.

To conclude, we conduct experiments on the CelebA-Dialogue dataset and evaluate the outcomes using the VisoGender benchmark. Results consistently reveal that targeted debiasing of individual encoders mitigates gender bias more effectively while preserving overall model performance. By demonstrating that targeted interventions reduce gender bias while preserving performance, our work contributes practical insights for building fairer vision-language systems.

Limitations

Despite these contributions, our study has limitations. First, the use of binary gender annotations excludes non-binary and LGBTQ+ identities, restricting the inclusiveness of our evaluation. Second, our focus is limited to gender bias and does not consider intersectional biases, such as those related to race or age.

Future Work

In future work, we plan to broaden the scope of our analysis to address intersectional biases, such as those involving race, age, and skin tone, which may interact with gender in complex ways. This would allow for a more nuanced understanding of model fairness across diverse identities. Additionally, investigating the temporal and contextual dynamics of bias—such as how models adapt to evolving cultural norms or contextual cues can offer deeper insights into the stability and robustness of debiasing methods.

Another important direction is exploring bias mitigation strategies during the pretraining phase, rather than only through fine-tuning, to assess whether early interventions result in more systemic improvements. Finally, we plan to test our methods in real-world deployment scenarios such as image captioning, content moderation, and recommendation systems, to evaluate both fairness and utility in applied settings.

References

Ali Abdollahi, Mahdi Ghaznavi, Mohammad Reza Karimi Nejad, Arash Mari Oriyad, Reza Abbasi, Ali Salesi, Melika Behjati, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. 2024. *GABInsight: Exploring Gender-Activity Binding Bias in Vision-Language Models*. IOS Press.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. 2023. Sentence-level prompts benefit composed image retrieval. *arXiv preprint arXiv:2310.05473*.

L. Darur, S.K. Gouravarapu, S. Goel, and P. Kumaraguru. 2024. Improving bias metrics in vision-language models by addressing inherent model disabilities. In *Workshop on Algorithmic Fairness through the Lens of Metrics and Evaluation, NeurIPS 2024*.

Omkar Dige, Diljot Arneja, Tsz Fung Yau, Qixuan Zhang, Mohammad Bolandraftar, Xiaodan Zhu, and Faiza Khan Khattak. 2024. Can machine unlearning reduce social bias in language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 954–969, Miami, Florida, US. Association for Computational Linguistics.

Daniel Fitousi. 2021. Stereotypical processing of emotional faces: Perceptual and decisional components. *Frontiers in Psychology*, 12.

Samarth Garg, Vivek Hruday Kavuri, Gargi Shroff, and Rahul Mishra. 2025. Ktcr: Improving implicit hate detection with knowledge transfer driven concept refinement.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic.

Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Diego A. B. Moreira, Alef Iury Ferreira, Jhessica Silva, Gabriel Oliveira dos Santos, Luiz Pereira, João Medrado Gondim, Gustavo Bonil, Helena Maia, Nádia da Silva, Simone Tiemi Hashiguti, Jefersson A. dos Santos, Helio Pedrini, and Sandra Avila. 2024. Fairpivara: Reducing and assessing biases in clip-based multimodal models.

Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R. Varshney. 2018. Understanding unequal gender classification accuracy from face images.

Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Improving generalizability in implicitly abusive language detection with concept activation vectors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5517–5529, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Tejas Srinivasan and Yonatan Bisk. 2022. Worst of both worlds: Biases compound in pre-trained vision-and-language models.

Ryan Steed and Aylin Caliskan. 2021. 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, page 701–713. ACM.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance ReWriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, Florence, Italy. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models.

Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. 2024. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective.

Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2023. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *2023 ACM Conference on Fairness Accountability and Transparency*, FAccT '23, page 1174–1185. ACM.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Yisong Xiao, Aishan Liu, QianJia Cheng, Zhenfei Yin, Siyuan Liang, Jiapeng Li, Jing Shao, Xianglong Liu, and Dacheng Tao. 2024. Genderbias-*VL*: Benchmarking gender bias in vision language models via counterfactual probing.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2022. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. 2023. Composing parameter-efficient modules with arithmetic operations.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2020. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology.