Images Speak Louder than Words: Understanding and Mitigating Bias in Vision-Language Model from a Causal Mediation Perspective

Anonymous ACL submission

Abstract

Vision-language models (VLMs) pre-trained on extensive datasets can inadvertently learn biases by correlating gender information with specific objects or scenarios. Current methods, which focus on modifying inputs and monitoring changes in the model's output probability scores, often struggle to comprehensively understand bias from the perspective of model components. We propose a framework that incorporates causal mediation analysis to measure and map the pathways of bias generation and propagation within VLMs. This approach allows us to identify the direct effects of interventions on model bias and the indirect effects of interventions on bias mediated through different model components. Our results show that image features are the primary contributors to bias, with significantly higher impacts than text features, specifically accounting for 32.57% and 12.63% of the bias in the MSCOCO and PASCAL-SENTENCE datasets, respectively. Notably, the image encoder's contribution surpasses that of the text encoder and the deep fusion encoder. Further experimentation confirms that contributions from both language and vision modalities are aligned and non-conflicting. Consequently, focusing on blurring gender representations within the image encoder which contributes most to the model bias, reduces bias efficiently by 22.03% and 9.04% in the MSCOCO and PASCAL-SENTENCE datasets, respectively, with minimal performance loss or increased computational demands.

1 Introduction

Vision-language models have shown promising results in tasks such as classification (Li et al., 2023; Jia et al., 2021; Radford et al., 2021), image search (Sun et al., 2023; Radford et al., 2021; Li et al., 2023), and object detection (Kuo et al., 2023; Li et al., 2022) by training on large-scale image-text pairs to understand the correspondences between cross-modal image features and language features. Models trained on extensive datasets exhibit excellent zero-shot capabilities (Radford et al., 2021; Yu et al., 2022; Li et al., 2022; Zhang et al., 2022a) but also risk discovering and exploiting societal biases present in the underlying image-text pair corpora, potentially introducing bias that leads to social unfairness (Zhao et al., 2017). The revelation, measurement, and understanding of biases within models (Zhou et al., 2022; Zhang et al., 2022b; Lee et al., 2023; Vig et al., 2020) has sparked widespread research interest and are crucial for bias mitigation (Zhang et al., 2022b; Seth et al., 2023; Dehdashtian et al., 2023). However, most contemporary methods, derived from language models, lack standardized metrics for evaluating bias and primarily assess the correlation between the outputs of classifiers and external attributes (Zhang et al., 2022b). Barrett et al. (2019) noted that interpretations based on classifier outputs can be factually inaccurate and not generalizable. While these methods can highlight the impacts of certain contributions on model outputs, they (1) fail to comprehend the generation and flow of bias within the model and (2) do not understand the causal roles of model components in the generation and propagation of bias. Consequently, they are not able to provide clear guidance on how to effectively mitigate bias at the model level.

In this work, we propose a standardized framework to measure bias in VLMs, providing a comprehensive understanding of how bias flows within the entire model structure. Specifically, we use the GLIP model (Li et al., 2022) as a case study, focusing on gender bias in the task of object detection, which is a predominant and challenging problem in computer vision. We conduct the analysis on both the MS-COCO (Lin et al., 2014) and PASCAL-SENTENCE (Rashtchian et al., 2010) datasets. We observe that GLIP model exhibits unbalanced inference capabilities on different genders, with certain indoor object categories like pets more likely to be associated with females and outdoor objects like vehicles with males. To holistically understand how the bias flows in the model, we adapt causal mediation analysis (Vig et al., 2020) to VLMs, providing a finer-grained study of contributions from different model components. We find that, among the different model components (text module, image module, and fusion module that combines them), the image module contributes the most to the model's bias – over twice as much as the text module. In the MSCOCO and PASCAL-SENTENCE datasets, image features accounted for 32.57% and 12.63% of the bias generated, respectively, compared to approximately 15.48% and 5.64% by text features. Also, the interaction and updating process between image and text features during the deep fusion process significantly impacts bias production, accounting for about 57% of the contributions in the image and text encoders. Furthermore, by integrating interventions across different modules, we discovered that their contributions to bias are aligned and do not conflict, allowing us to prioritize bias mitigation efforts within the image encoder, which is the most substantial contributor to bias. Based on the results, we propose to effectively mitigate the bias in VLMs: reducing biases from the image module can successfully reduce bias by 22.03% on the MSCOCO dataset and 9.04% on the PASCAL-SENTENCE dataset, compared to a reduction of 7.08% and 1.18% in the text module. In summary, the contributions of our work are:

- We provide a comprehensive evaluation of the bias in VLMs, with an understanding of the contribution from each model module, which is missing in the literature.
- We analyze the correlation between the biases from different modules and discover that the bias in different modules are aligned and do not conflict with each other.
- We propose an effective bias mitigation strategy to reduce the bias from the module that contributes most to the model bias when facing a limited budget.

2 Related Work

In recent years, vision-language models (VLMs) have experienced rapid advancements. The latest developments in VLMs often employ a dual-stream architecture that separately encodes text and images (Kim et al., 2021), and these are then merged and aligned to facilitate cross-modal understanding of visual and linguistic features (Radford et al., 2021). Furthermore, some studies treat the joint training of images and text as a phrase localization process, aiming to better align and integrate visual and linguistic features (Li et al., 2022). Typically, these models are trained on image-text pairs from datasets such as MSCOCO (Lin et al., 2014), VQA (Antol et al., 2015), OpenImages (Kuznetsova et al., 2020), and Flickr30k Entities (Plummer et al., 2015), achieving impressive results in various downstream tasks including image classification (Radford et al., 2021), image generation (Radford et al., 2021), visual question answering (Li et al., 2018; Antol et al., 2015), and image captioning (Lu et al., 2019; Alayrac et al., 2022).

Alongside their development, the societal biases exhibited by VLMs have also attracted significant attention. These models often reflect societal stereotypes and may even amplify such biases (Zhou et al., 2022). Most contemporary research addressing bias in VLMs has borrowed methodologies from language model studies. For instance, Srinivasan and Bisk (2021) utilized a language masking model to explore gender biases by using templates containing a specific entity and analyzing the probabilities of masked entities (Kurita et al., 2019). Some researchers have examined biases through the comparison of factual and counterfactual inputs, with Zhang et al. (2022b) investigating biases by examining predicted probabilities from both factual and counterfactual inputs, and Howard et al. (2024) using the Perspective API to score predictions derived from such inputs to study model biases.

However, existing evaluation methods primarily observe changes in the probability scores of model outputs following interventions on input samples. This approach limits our understanding of the underlying causes of bias generation and propagation within model components (Barrett et al., 2019). Therefore, we propose a standardized framework for evaluating bias in vision-language tasks and introduce causal mediation analysis (Robins and Greenland, 1992; Pearl, 2022; Vig et al., 2020) within the context of vision-language models. This methodology helps us understand the pathways of bias generation and propagation from the input level to model components.

3 Bias Measurement and Understanding in VLM

In this section, we propose a bias evaluation metric to assess the bias of VLM in the object detection task. By applying causal mediation analysis, we quantify the contribution on bias from various components within the model pipeline which helps us trace the origins and propagation of bias throughout the model pipeline. Additionally, we investigate the interactions between different modalities to understand how they collectively influence model bias which will be used as guidance for bias mitigation later.

3.1 Bias Evaluation Metrics

In the literature, there have been various methodologies proposed to measure bias, including notable contributions from Zhao et al. (2017), Wang and Russakovsky (2021) and Zhao et al. (2023). These studies often assess bias amplification by comparing statistics between the training dataset and model outputs, where the models are trained and tested on similarly distributed datasets. In contemporary settings, most VLMs undergo training on extensive collections of image and text corpora. In real-world applications, users may fine-tune a model on a dataset specific to a downstream task. The combination of fine-tuning data and pre-training data can introduce noise, complicating the statistics of previously mentioned bias evaluations. Additionally, many pre-training datasets used for large-scale models are either difficult to access or require significant computational resources for analysis, making existing evaluations challenging to deploy in modern settings.

Notably, recent advancements in VLM have demonstrated impressive zero-shot performance, enabling models to make accurate predictions on benchmark datasets without any fine-tuning (Radford et al., 2021; Yu et al., 2022; Li et al., 2022; Zhang et al., 2022a). In our study, we explore a zero-shot scenario where VLMs are directly tasked with making predictions on a benchmark dataset without any fine-tuning.

Drawing inspiration from observations in Zhao et al. (2017), where females typically correlate more closely with indoor objects than males, we introduce the definition of $\textsc{Bias}_{\text{VL}}$ which captures model's underlying correlations between sensitive attributes (e.g., genders) and objects:

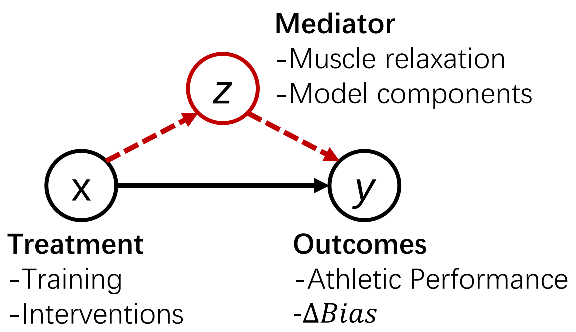


Figure 1: Causal Mediation Analysis example

$$\textsc{Bias}_{\text{VL}} := \sum_{\text{object}} |\mathcal{C}(\text{object, male}) \\ - \mathcal{C}(\text{object, female})| \quad (1)$$

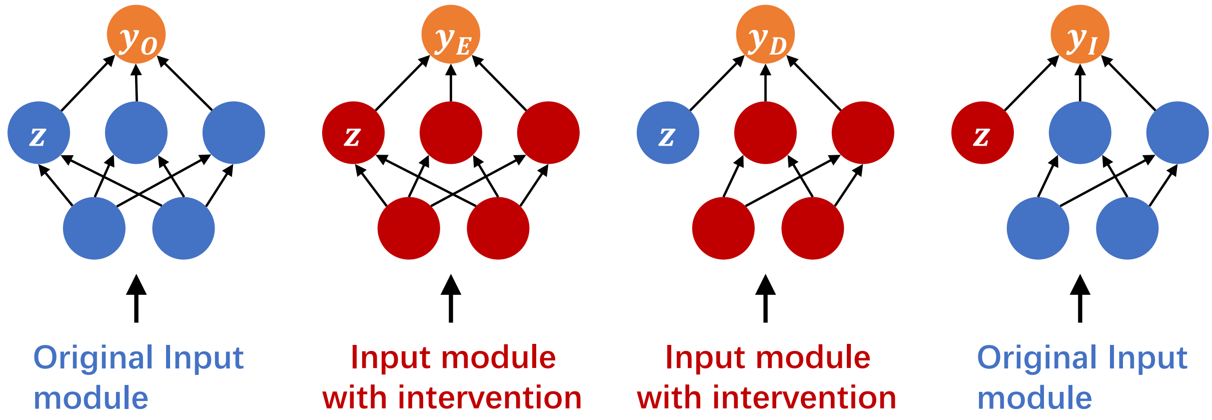
where $\mathcal{C}(x, y)$ measures the correlation between x and y. In our case, we use a false positive rate (FPR) to describe the correlation, which measures how often one specific gender y can trigger a model to incorrectly predict one object x in the image. [1]

3.2 Causal Mediation Analysis Method

Causal mediation analysis measures how a treatment effect influences an outcome either directly or indirectly through a mediator variable (Robins and Greenland, 1992; Vig et al., 2020; Robins, 2003; Pearl, 2022). An illustrative example is shown in Figure 1, where athletes engage in strength training to improve athletic performance. After training, they need muscle relaxation to alleviate soreness, which also impacts performance. Thus, strength training can have a direct effect on athletic performance through its intended mechanisms and an indirect effect through muscle relaxation.

In our study, the treatment consists of interventions on the input module, while the mediator could be any model component or finer-grained layer or neuron we are interested in and the outcome is the change in gender bias in the model's prediction results. Therefore, we define three types of intervention: a) `replace-gender`, which replaces the gender word *man* or *woman* to a gender-neutral word *person* in the text of the input module; b) `mask-gender`, where pixels corresponding to a person in the image module are masked, thus removing gender information from the input images; and c) `null`, which leaves the original text and image modules unchanged.

[1]Following existing work, we also consider binary gender



(a) Baseline

(b) Intervention Effect $E := y_E - y_O$

(c) Direct Effect $DE := y_D - y_O$

(d) Indirect Effect $IE := y_I - y_O$

Figure 2: Bias understanding with causal mediation analysis. In the diagram, z represents the mediator, and y_O, y_E, y_D, y_I represent the bias values of the model's output under various interventions. The ***intervention effect*** quantifies the change in the bias score under the specified intervention; the ***direct effect*** quantifies the change in bias score resulting from an intervention in the input module while maintaining the mediator in the state of a `null` intervention; the ***indirect effect*** measures the change in the bias score when the input module remains unchanged, but the mediator is set to the state of a specific intervention.

We perform causal mediation analysis on the GLIP model by introducing interventions in the input module and observing changes in BIAS_{VL} values defined in Eq.(1). Following Vig et al. (2020), we define the ***Direct Effect (DE)*** as changes in the BIAS_{VL} score when the intervention is applied to the input module while the mediator (model components) remains in the '`null`' state of intervention. The ***Indirect Effect (IE)*** represents changes in the bias score when the input module is fixed, but the mediator is set in the state of a certain intervention. We can select any model structure of interest as the mediator and choose '`mask-gender`', '`replace-gender`', or combinations of them as interventions in the input module (Figure 2).

4 Experimental Setup of Bias Measurement and Understanding

Model For the object detection task, we employed the GLIP model pre-trained on the O365, GoldG, CC3M, and SBU datasets (Li et al., 2022). The model consists of an image module, a text module, and a deep-fusion module that updates and aligns image features and text features. For object detection, the GLIP model makes predictions based on the given image and a text input, which is a list of possible categories separated by commas.

Dataset Our experiments were conducted on the MSCOCO and PASCAL-SENTENCE datasets. For MSCOCO, we follow the setting in Zhao et al. (2017), where we only consider 66 objects that appear with man or woman more than 100 times in the training data. For the PASCAL-SENTENCE dataset, which includes 20 categories but lacks gender labels, we annotated gender based on the five captions associated with each image. An image is labeled as male if any caption mentions "male, males, man, men, boy, boys" and as female if any caption mentions "female, females, woman, women, girl, girls". Images that do not include any person or mention both genders were excluded.

Interventions on image encoder and text encoder Initially, we implement `replace-gender` and `mask-gender` interventions on the inputs respectively without any alterations to the model components. By monitoring the changes in the values of BIAS_{VL}, the individual impacts of image and text inputs on gender bias within the input module were assessed. Subsequently, we conducted a detailed causal mediation analysis on the text encoder and image encoder, respectively, by choosing the attention head within a specific layer and those in all preceding layers as mediators, conducting experiments from shallow to deep layers. This analysis aimed to identify whether the text encoder or image encoder contributes more significantly to

in this study.

gender bias and to determine which layers in the model are principally responsible for bias generation. It also sought to understand how bias flows and accumulates across different layers within the encoders. Then, we selected a combination of attention layers from both the image encoder and text encoder as mediators to observe changes in bias and compare these results with previous findings, exploring whether different modalities reinforce bias or conflict in the direction of bias.

Interventions on deep-fusion encoder In the deep fusion encoder, where image and text features dynamically interact and are updated, we implement `replace-gender` and `mask-gender` interventions in the input module to control the state of image and text features within the deep fusion module. We also select the attention heads within a specific layer and all preceding layers' attention heads as the mediator for conducting causal mediation analysis. By observing changes in the values of BIAS_{VL}, we explore how image and text features individually affect the deep fusion process and subsequently influence bias generation.

5 Results

5.1 Bias Measurement

We present the results of BIAS_{VL} in Table 1, for the MSCOCO dataset, without any intervention on the inputs, the BIAS_{VL} measured was 1.434. To highlight the significance of this bias, we randomly divided subsets composed of male images into two equal parts, achieving an BIAS_{VL} of 0.278. Similarly, dividing female image subsets randomly resulted in an BIAS_{VL} of 0.359. Both results are significantly lower than 1.434, and comparable results were observed with the PASCAL-SENTENCE dataset, as detailed in Table 1. The results in the random division demonstrate that a model with balanced inference capabilities across a dataset would yield minimal BIAS_{VL} values when divided into equal subsets (i.e., the gender stays the same). However, when model predictions are influenced by attributes such as gender, splitting the dataset based on such attributes leads to higher BIAS_{VL} values.

We also provide detailed statistics of False Positive Rate (FPR) scores for various objects in the PASCAL-SENTENCE dataset, presented in Figure 3. Our statistics reveal that a significant portion of indoor objects, such as furniture and pets, exhibit higher FPRs in images of females than in those of males. Conversely, outdoor objects, such as vehicles, tend to have higher misclassification rates in images of males. These findings suggest that the model more closely associates females with indoor objects. The FPR scores for different objects on the MSCOCO dataset are included in the appendix.

Dataset	BIAS_{VL}	$\text{BIAS}_{\text{VL}}(\text{M}_1,\text{M}_2)$	$\text{BIAS}_{\text{VL}}(\text{F}_1,\text{F}_2)$
MSCOCO	1.434	0.278	0.359
PASCAL-S	1.369	0.341	0.381

Table 1: BIAS_{VL} for MSCOCO and PASCAL-SENTENCE (PASCAL-S) Datasets without any intervention. M and F stand for "male" and "female" respectively. BIAS_{VL} values obtained in two sets of images with the same gender are significantly lower than the BIAS_{VL} obtained from datasets divided by gender.

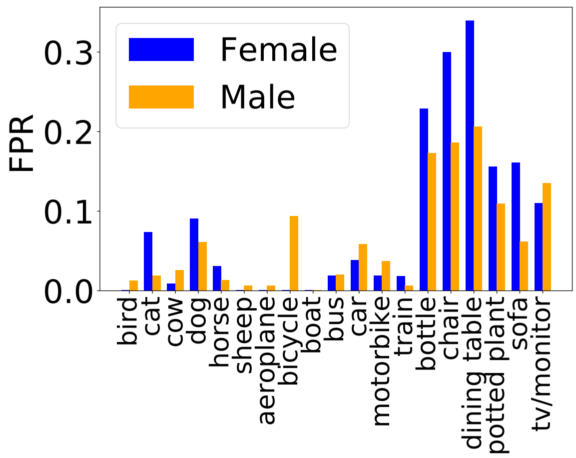
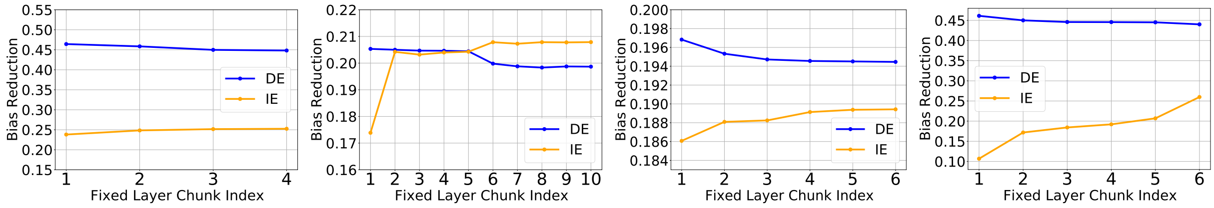


Figure 3: False Positive Rate (FPR) for various objects in the PASCAL-SENTENCE dataset. For most pets and indoor objects, the FPR is higher in images of females than in those of males; conversely, for most outdoor objects such as vehicles, the FPR is higher in images of males than in those of females. These results indicate that females correlate more closely with indoor objects than males.

5.2 Bias Understanding with Causal Mediation Analysis

We conduct the causal mediation analysis on different modules to study their effect on the model bias. We find that the image module influences the model bias more than the text module and the fusion module. In addition, we show that the bias in the image and text modules are aligned – they are showing similar gender bias tendencies rather than conflicting ones.

Image encoder Applying the `mask-gender` intervention to the input image module reduced the

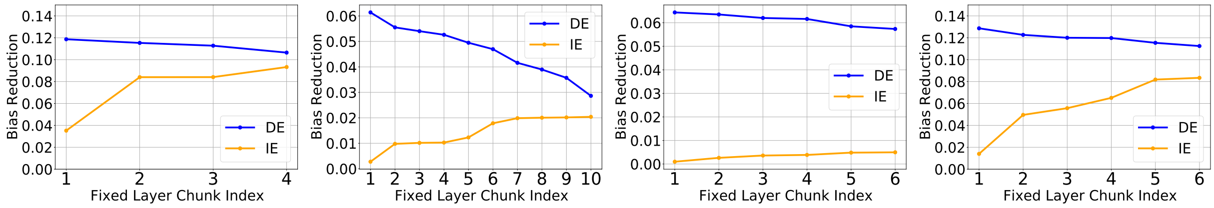


(a) image encoder (COCO)

(b) text encoder (COCO)

(c) text part of deep-fusion encoder (COCO)

(d) image part of deep-fusion encoder (COCO)



(e) image encoder (PASCAL-S)

(f) text encoder (PASCAL-S)

(g) text part of deep-fusion encoder (PASCAL-S)

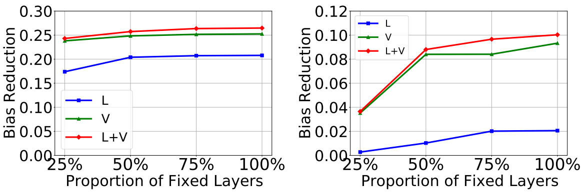
(h) image part of deep-fusion encoder (PASCAL-S)

Figure 4: Causal mediation analysis of bias on the COCO and PASCAL-S (PASCAL-SENTENCE) datasets. Panels (a) and (e) show the DE (Direct Effect) and IE (Indirect Effect) for the image module; Panels (b) and (f) represent the DE and IE for the text module; Panels (c) and (g) illustrate the DE and IE for the text part of the deep-fusion encoder, and panels (d) and (h) for the image part of the deep-fusion encoder. The findings highlight that image features contribute more significantly to bias than text features, with the image module being the primary contributor to model bias.

BIAS$_{\text{VL}}$ to 0.967 for the MSCOCO dataset and to 0.664 for the PASCAL-SENTENCE dataset, representing reductions of approximately 32.57% and 12.63%, respectively. We employed the attention heads in the image encoder as the mediator to examine both the indirect effects of this model component and the direct effects of the `mask-gender` on predictions. Figure 4a and Figure 4e illustrate that employing more attention heads as mediators leads to greater reductions in indirect effect, with diminishing reductions in direct effect. This supports an intuition that removing gender information from more layers in the image encoder weakens the model's dependency on latent correlations between gender in images and specific objects, thus mitigating gender bias in predictions. Furthermore, while interventions at the input level significantly impact final predictions, targeting the image encoder alone achieves about 53% of the `mask-gender` effect.

Text encoder Implementing a `replace-gender` intervention on the input text module reduced the BIAS$_{\text{VL}}$ to 1.212 for the MSCOCO dataset and to 0.720 for the PASCAL-SENTENCE dataset, reductions of approximately 15.48% and 5.64%, respectively. We chose the attention heads within the text encoder as the mediator in this case. As shown in Figure 4b and Figure 4f, similar to the image encoder insights, removing gender information from multiple layers in the text encoder substantially decreases the model's reliance on latent correlations between gender in text and specific objects, thereby reducing prediction biases. The `replace-gender` intervention led to a smaller reduction in bias compared to `mask-gender`, emphasizing the more substantial role of images in generating gender bias relative to text. This outcome is likely influenced by the simplistic structure of the input text used in our study, which adheres to the format described in original GLIP experiments (Li et al., 2022), separating each category with a period, resulting in less complex text features than image features. Language models typically capture basic features such as syntactic structures at shallow layers and more complex semantic information at deeper layers, correlating with the significant changes in BIAS$_{\text{VL}}$ observed at the sixth layer.

Deep fusion encoder To further validate whether image features contribute more to bias creation than text features, we utilized the attention heads in the deep fusion encoder as the mediator, adjusting the attention heads' parameters in the states of either `mask-gender` intervention or `replace-gender` intervention. The results displayed in Figure 4d and Figure 4c show that for the MSCOCO dataset, the indirect effects from `mask-gender` and `replace-gender` through the deep fusion encoder are up to 0.260 and 0.189, respectively, reducing the $\textsc{Bias}_{\text{VL}}$ by approximately 18.13% and 13.18% (Figure 4h and Figure 4g). For the PASCAL-Sentence dataset, the reductions are 10.80% and 0.53%, respectively. These findings reaffirm our conclusion that image features play a more substantial role in bias generation than text features. They also suggest that even though the deep fusion module does not extract features directly from images and text, the interactive updating process between text and image features significantly influences bias generation, accounting for approximately 55.70% of the effect observed with the encoder alone.



(a) MSCOCO (b) PASCAL-SENTENCE

Figure 5: Comparison of Bias Reduction Across Modalities with Interventions in Vision (V) and Language Modules (L) on the MSCOCO and PASCAL-SENTENCE datasets. **V** represents the results of interventions in the vision modality, **L** represents the results of interventions in the language modality, and **L+V** represents the results of simultaneous interventions in both the vision and language modalities. The contributions to bias from the two modalities are aligned and non-conflicting. Intervening simultaneously in both the visual and language modalities results in a greater reduction of bias compared to interventions in any single modality alone.

Interventions comparison Multi-modal models consist of various interacting modules, each of which can learn distinct biases. However, the current literature does not thoroughly investigate whether these biases are aligned or disparate across different modules. In this section, we conduct an empirical analysis in VLMs to address this question. We simultaneously intervene in both the vision and language modalities. We apply `replace-gender` and `mask-gender` interventions to the input module and select a consistent proportion of attention heads in both the image encoder and text encoder as mediators. This setup allows us to observe changes in $\textsc{Bias}_{\text{VL}}$ and compare these with the changes induced by interventions in single modalities. Figure 5a and Figure 5b demonstrate that combined interventions on both images and text achieve greater bias reduction than interventions on either alone. However, the total reduction is not merely additive; the overall bias reduction is less than the sum of the individual contributions.

6 Bias Mitigation Method

Based on our experimental results, image features contribute most significantly to gender bias and the image encoder has a more pronounced impact on bias compared to the text encoder and deep-fusion encoder. Therefore, our intuition is that focusing on reducing gender representation in the image encoder will effectively reduce bias, especially when facing a computation budget. We use the bias mitigation achieved from the text encoder as a baseline, then focus on reducing bias from the image encoder and compare the results with the baseline.

Text Encoder For the text encoder, we aim to blur the gender representation in text features. We modify the structure of the text encoder to first identify gender-related terms *(man, woman, men, women, male, female, males, females)* in the input text. A new sentence is generated by replacing these gendered terms with their corresponding anti-gender terms (i.e., *man* to *woman*, *male* to *female*). The text encoder's output features are the average of the original sentence's text features and the anti-gender sentence's text features. Since the only difference between the two sentences is the gendered terms, this approach effectively blurs gender representation within the text encoder. We then let model to make predictions and observe the reduction in $\textsc{Bias}_{\text{VL}}$.

Image Encoder Similarly, for the image encoder, we aim to blur gender representation in image features. To achieve this, we incorporate MTCNN (Zhang et al., 2016) as a face detector and MobileNet (Sandler et al., 2018) as a gender classifier into the existing image encoder framework. Both networks are lightweight, allowing their integration without significantly increasing the computational load during inference. When an image is input into the image encoder, the MTCNN (Zhang et al., 2016) network first identifies potential faces and outlines them with bounding boxes. MobileNet (Sandler et al., 2018) then classifies the gender of the faces within these boxes.

We have prepared a male face image and a female face image in advance. Depending on the gender predicted by MobileNet (Sandler et al., 2018), we replace the face in the bounding box with the corresponding pre-prepared anti-gender face im-

	AP		Bias		Bias Mitigated	
	MSCOCO	PASCAL-S	MSCOCO	PASCAL-S	MSCOCO	PASCAL-S
GLIP-T	46.6	68.4	1.434	0.763	0	0
GLIP_ImageFair	46.2	68.3	1.118	0.694	22.03%	9.04%
GLIP_TextFair	46.6	68.4	1.322	0.754	7.8%	1.18%

Table 2: Performance comparison of different methods on MSCOCO and PASCAL-S (PASCAL-SENTENCE) datasets. AP (Average Precision) is the metric used for zero-shot object detection. "`GLIP`" represents the original GLIP model, "`GLIP_ImageFair`" denotes the model with bias mitigation implemented in the image encoder, and "`GLIP_TextFair`" refers to the model with bias mitigation applied in the text encoder. Intervention in the image encoder is more effective than the text encoder in reducing the bias score without significant performance loss.

age. The final image features output by the image encoder are an average of the original image features and the features of the newly introduced anti-gender face. This method effectively blurs the original gender representation in the image. Then we let the model to make predictions and observe the reduction in BIAS_{VL}.

7 Experimental Setup of Bias Mitigation

Model We utilized the GLIP model, pre-trained on the O365, GoldG, CC3M, and SBU datasets (Li et al., 2022). In our setup, we incorporated an MTCNN (Zhang et al., 2016) pre-trained on the Wider Face and CelebA datasets as a face detector within the image encoder. Additionally, we integrated a MobileNet (Sandler et al., 2018) pre-trained on ImageNet to serve as a gender classifier.

Dataset We evaluated the effectiveness of bias mitigation on the MSCOCO and PASCAL-SENTENCE datasets. To assess the model's object detection performance, we compared it with the original GLIP (Li et al., 2022) on the MSCOCO and PASCAL-SENTENCE datasets using the AP (Average Precision) metric for zero-shot object detection.

8 Results

As indicated in Table 2, blurring gender representations in the image encoder demonstrated significant bias mitigation on both the MSCOCO and PASCAL-SENTENCE datasets. The experimental findings suggest that obscuring gender information in the image encoder is more effective at reducing model bias compared to similar interventions in the text encoder. Our results show that by blurring gender representations in the image features within the image encoder, we effectively reduced model bias by 22.03% and 9.04% on the MSCOCO and PASCAL-SENTENCE datasets, respectively, with minimal impact on model performance.

9 Conclusion

Vision-language models (VLMs) trained on large-scale image-text pair corpora are at risk of learning social biases from their training data. In this paper, we introduced a standardized framework incorporating causal mediation analysis to measure and understand the pathways through which model bias is generated and propagated within VLMs. We discovered that image features contribute significantly more to model bias than text features, and the contributions from the image encoder substantially exceed those from the text encoder and deep fusion encoder. Furthermore, the contributions to bias from different language modalities reinforce each other. Subsequently, by focusing on the components that contribute most to bias, we efficiently reduced model bias.

Our work provides a framework for measuring, understanding, and mitigating model bias, which, although utilized here within the realm of object detection, can be extended to a wide range of VLM tasks. However, our framework is primarily applicable to white-box models, as it requires interventions at the internal components of the model. A promising direction for future work would involve expanding our framework to encompass additional modalities such as audio or video. This expansion could further enhance our understanding of multimodal interactions and their impact on bias, as well as deepen insights into how different sensory inputs contribute to, or mitigate, biases in AI systems.

10 Limitations

Our work provides a framework for measuring, understanding, and mitigating model bias in vision-language models (VLMs), with broad applicability across various VLM tasks. However, our approach primarily applies to white-box models, as it requires interventions within the model's internal components. Consequently, this limitation implies that our methods might not be directly applicable to scenarios where model internals are inaccessible or when dealing with black-box systems.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335.

Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. 2023. Fairvlm: Mitigating bias in pre-trained vision-language models. In *The Twelfth International Conference on Learning Representations*.

Phillip Howard, Anahita Bhiwandiwalla, Kathleen C Fraser, and Svetlana Kiritchenko. 2024. Uncovering bias in large vision-language models with counterfactuals. *arXiv preprint arXiv:2404.00166*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. 2023. F-vlm: Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations (ICLR)*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Nayeon Lee, Yejin Bang, Holy Lovenia, Samuel Cahyawijaya, Wenliang Dai, and Pascale Fung. 2023. Survey of social bias in vision-language models. *arXiv preprint arXiv:2309.14381*.

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.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.

Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 139–147.

James M Robins. 2003. Semantics of causal dag models and the identification of direct and indirect effects. *Highly structured stochastic systems*, pages 70–82.

James M Robins and Sander Greenland. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829.

Tejas Srinivasan and Yonatan Bisk. 2021. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *arXiv preprint arXiv:2104.08666*.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2023. Alpha-clip: A clip model focusing on wherever you want. *arXiv preprint arXiv:2312.03818*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In *International Conference on Machine Learning*, pages 10882–10893. PMLR.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. 2022a. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Yi Zhang, Junyang Wang, and Jitao Sang. 2022b. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4996–5004.

Dora Zhao, Jerone Andrews, and Alice Xiang. 2023. Men also do laundry: Multi-attribute bias amplification. In *International Conference on Machine Learning*, pages 42000–42017. PMLR.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Kankan Zhou, Yibin LAI, and Jing Jiang. 2022. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. Association for Computational Linguistics.

A Appendix

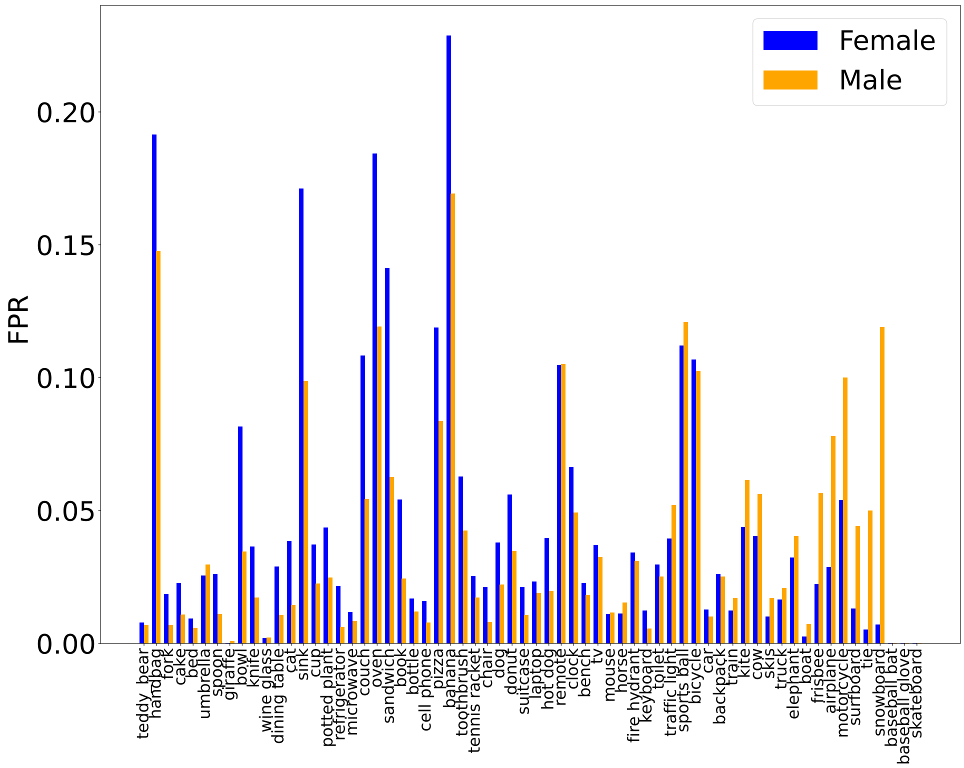


Figure 6: False Positive Rate (FPR) for various objects in the MSCOCO dataset. For most indoor objects, the FPR is higher in images of females than in those of males; conversely, for most outdoor objects such as vehicles, the FPR is higher in images of males than in those of females. These results indicate that females correlate more closely with indoor objects than males.