Masking Latent Gender Knowledge for Debiasing Image Captioning

Fan Yang, Shalini Ghosh, Kechen Qin, Prashan Wanigasekara,
Emre Barut, Chengwei Su, Rahul Gupta, Weitong Ruan
Amazon AGI, MA, USA
{fyaamz, ghoshsha, qinkeche, wprasha, ebarut, chengwes, gupra, weiton}@amazon.com

Abstract

Large language models incorporate world knowledge and present breakthrough performances on zero-shot learning. However, these models capture societal bias (e.g., gender or racial bias) due to bias during the training process which raises ethical concerns or can even be potentially harmful. The issue is more pronounced in multi-modal settings, such as image captioning, as images can also add onto biases (e.g., due to historical non-equal representation of genders in different occupations). In this study, we investigate the removal of potentially problematic knowledge from multi-modal models used for image captioning. We relax the gender bias issue in captioning models by degenderizing generated captions through the use of a simple linear mask, trained via adversarial training. Our proposal makes no assumption on the architecture of the model and freezes the model weights during the procedure, which also enables the mask to be turned off. We conduct experiments on COCO caption datasets using our masking solution. The results suggest that the proposed mechanism can effectively mask the targeted biased knowledge, by replacing more than 99% gender words with neutral ones, and maintain a comparable captioning quality performance with minimal (e.g., -1.4 on BLEU4 and ROUGE) impact to accuracy metrics.

1 Introduction

Large models are known to have harmful biases. One example is gender bias, where the model learns incorrect correlation between gender and objects, occupations, etc. As these result from inherent bias presented in the data, this process is almost impossible to govern – especially considering the scale of data required for training these models. In addition, recent works have shown that these models can exacerbate such biases from the training data at test time (Hendricks et al., 2018; Wang and Russakovsky, 2021).

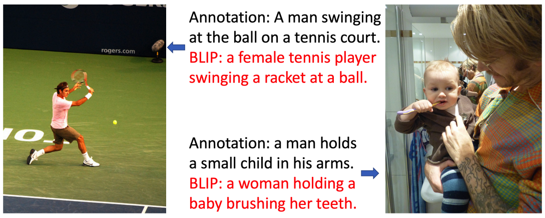

Annotation: A man swinging at the ball on a tennis court.
BLIP: a female tennis player swinging a racket at a ball.
Annotation: a man holds a small child in his arms.
BLIP: a woman holding a baby brushing her teeth.

Figure 1: BLIP model mis-classifies gender when generating captions.

Due to the training cost of large models, it is often difficult to address such model vulnerabilities by re-training. Some recent works propose to locate a subset of model parameters that cause issues and subsequently edit them (Santurkar et al., 2021; Jang et al., 2022; Mitchell et al., 2022b), while others propose to use prompting with in-context examples (Murty et al., 2022) and meta-learning to prevent large models from learning harmful biases (Mitchell et al., 2022a). While these works mostly focus on text-based models, the computer vision community has also been fighting undesirable biases in visual question answering (Hirota et al., 2022a), image captioning (Zhao et al., 2017; Hendricks et al., 2018; Zhao et al., 2021; Tang et al., 2020), and image classification (Yao et al., 2022; Wang et al., 2022).

In this work, we study how to debias image captioning models with respect to the gender attribute. Studies have shown that generated descriptions can refer to an incorrect gender, e.g., identify a woman riding motorcycle as a man and a man in a kitchen as a woman. We illustrate the problem using the state-of-the-art captioning model BLIP (Li et al., 2022b) in Figure 1. Image captioning models often rely on an encoder-decoder framework, which encodes raw images to continuous representations and the decoder generates the captions autoregressively. State-of-the-art methods, such as BLIP (Li et al., 2022b), BLIP-2 (Li et al., 2023), and LLaVA (Liu et al., 2023b,a), leverage

Proceedings of the 4th Workshop on Trustworthy NLP (TrustNLP 2024), pages 227–238
June 21, 2024 ©2024 Association for Computational Linguistics

pre-trained vision transformer and pre-trained language model to boost the performance. However, they also inherit some shortcomings of these methods: (i) there are no means to control the inherent data bias due to the size of training data; (ii) it is difficult to update the entire model due to re-training cost. Therefore, existing works on debiasing image captioning are limited because they require to re-train the model with an improved neural architecture (Hendricks et al., 2018; Tang et al., 2020).

Furthermore, the use of explicit gendered words in the captions may exclude individuals identifying as any of the non-binary gender groups. We posit that these biases can be mitigated if a captioning model outputs gender-neutral tokens such as "human" or "person" instead of "man" or "woman". In that aim, we consider generating de-genderized captions as a new direction to debias image captioning.

We deliver the above via a masking framework, where the image embeddings are transformed before they are ingested by the encoder/decoder components of a multi-modal model stack. The mask acts as a de-biasing filter that removes the gender relevant information in the embedding (ideally) without other loss of information. The mask only works with the deep image representation, and we argue that the downstream text decoder would generate de-genderized caption if the input is not revealing gender.

The main contribution of this work are:

- We propose an easy-to-implement solution to hide gender knowledge from image representations through training a low parameter model, a mask, and consequently achieve unbiased image captioning. To effectively train the mask, we leverage domain adversarial training (Ganin et al., 2015) and design negative log-likelihood loss to be maximized on gender words and minimized on other words.
- We conduct extensive experiments for ablation studies on variations of our implementation. We experiment with COCO Caption datasets (Lin et al., 2014), and present both quantitative and qualitative analyses. We show that the proposed method can replace more than 99% gender words with neutral ones.

2 Related Work

Model Debiasing in Language Models. Language models capture social biases from the data they are trained; presence of gender bias (Zhao et al., 2019; Bordia and Bowman, 2019; Dinan et al., 2020; Sun et al., 2019; Basta and Costa-jussà, 2021; Pessach and Shmueli, 2022; Kotek et al., 2023) and racial bias (Garg et al., 2018; Davidson et al., 2019; Gehman et al., 2020; Manzini et al., 2019; Mehrabi et al., 2021) in language models have been well documented. To mitigate the bias, a commonly employed data-driven technique called Counterfactual data augmentation (CDA) proposes to swap bias attribute words in a dataset to re-balance a corpus (Zmigrod et al., 2019; Dinan et al., 2020; Webster et al., 2020; Barikeri et al., 2021). The re-balanced corpus is then used for further training to debias a model. This method requires domain knowledge or human intervention to generate plausible counterfactuals and may introduce noise or inconsistency into the data (Lauscher et al., 2021; Qiang et al., 2022; Meade et al., 2022). Bolukbasi et al. (2016) study the use of orthogonal projection for eliminating gender biases in word embeddings, which was subsequently extended by Liang et al. (2020) to include debiasing of sentence embeddings. Other methods include using dropout regularization as a bias mitigation technique (Webster et al., 2020), discouraging the model from generating biased text by tuning prompt (Schick et al., 2021), or projecting the neural representations to a null-space of classifiers that are used to predict unwanted information (Ravfogel et al., 2020). Recently, the remarkable performance of large language models across various tasks has also brought significant attention to the biases they exhibit (Brown et al., 2020a; Basta and Costa-jussà, 2021; Liu et al., 2022; Guo et al., 2022; Zhuo et al., 2023).

Model Debiasing in Vision-language Models. Research on debiasing vision-language models can be categorized into three groups: (i) dataset-level debiasing that seeks to balance imbalanced data (Zhao et al., 2021), (ii) model-level debiasing that mitigates bias by adjusting the model structure (Hendricks et al., 2018; Tang et al., 2020), and (iii) prompt-level debiasing that utilizes prompts to measure and eliminate biases (Chuang et al., 2023). In the context of vision-language models trained via contrastive loss, there has been active research to debias the CLIP model (Radford et al., 2021). The authors of the original CLIP paper investigated the presence of bias within their own paper (Agarwal et al., 2021). Wang et al. (2021) suggest the removal of dimensions in the CLIP embedding that exhibit a strong correlation with gender attributes. Berg et al. (2022) demonstrate that incorporating learned embeddings at the beginning of text queries in CLIP models results in a reduction of multiple measures of bias.

3 Gender Knowledge Masking

In this section, we describe how to mask gender knowledge in a pre-trained image captioning model using a trained mask. We utilize the BLIP model (Li et al., 2022b) in our presentation and experiments but note that the method can be applied to any other similar architecture where a multi-modal encoder ingests image embeddings (e.g., ALBEF (Li et al., 2021), BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023b)).

3.1 Masking Embeddings

At a high level, we transform image embeddings from image encoder, e.g., ViT (Kolesnikov et al., 2021), to gender-neutral embeddings via a mask and linear transformation before feeding them to the text-decoder. The parameters of the mask are learned via adversarial training on gender-specific words while the model's other parameters are frozen. We provide details below.

Image Embedding Mask. The text-decoder stack ingests the images via image embeddings produced by the vision transformer. Instead of using the stack of visual embeddings, $\mathbf{e}^v$, we provide the text-decoder with new embeddings $\hat{\mathbf{e}}^v$, where each token in $\mathbf{e}^v$ goes through a learned affine transformation, $\boldsymbol{\theta} \in \mathbb{R}^{K \times K}$ where K is the size of each image embedding token. Specifically we provide the text decoder with $\hat{\mathbf{e}}^v$, where,

$$\begin{aligned}\hat{\mathbf{e}}^v &= [\hat{e}^v_{CLS}, \hat{e}^v_1, \ldots, \hat{e}^v_L] \\ &= [\boldsymbol{\theta} e^v_{CLS}, \boldsymbol{\theta} e^v_1, \ldots, \boldsymbol{\theta} e^v_L]\end{aligned}$$

We apply mask on image representation $\mathbf{e}^v$ rather than the internal embeddings of the text decoder or directly the raw image input due to following considerations: 1. Applying mask inside the text decoder brings more risk on degrading text generation, as the language modeling task is often less stable than representation learning, for which the image embeddings were trained for; 2. Masking the raw images is a far harder task. It does not prevent leaking gender bias: training datasets can rely on learned biased gender-object correlations; also it is not clear what gender distribution exists in the pre-training dataset (thus, directional bias amplification leaks (Wang and Russakovsky, 2021)) and even a balanced dataset could amplify the association between objects and gender (Wang et al., 2018).

Training the Mask. To train $\boldsymbol{\theta}$, we freeze all of the BLIP model weights, and optimize solely over $\boldsymbol{\theta}$ by minimizing the standard negative log-likelihood (NNL) loss function used for captioning:

$$\min_{\boldsymbol{\theta}} L = -\frac{1}{T} \sum_{i}^{T} \log p(y_t | y_1, y_2, \ldots y_{t-1}, I(\boldsymbol{\theta})) \tag{1}$$

where $p(\cdot)$ represents the text-decoder, y_t are the tokens in the caption, T is the number of tokens in the caption, and $I(\boldsymbol{\theta})$ is the information provided via the image embeddings through the cross-attention layers.

Adversarial Training. During training, we also leverage domain adversarial training (Ganin et al., 2015). Specifically, if the caption contains any gender words, the gradient for the loss of that token is reversed, and are combined with gradients with non-genderized replacements. The masking and gradient reverse can be achieved via a few lines of code, which we illustrate in Algorithm 1, where λ is a hyper-parameter used to control the magnitude of gradient.

Algorithm 1: Training procedure for the Mask in pseudo-code

```
Gradient Reverse
class GradReverse:
# FORWARD PASS: Do nothing
# BACKWARD PASS:
def backward(grad, $\lambda$, **kwargs):
    return grad.neg() * $\lambda$
Masking
$\mathbf{e}^v$ = VISUAL ENCODER(raw image)
$\hat{\mathbf{e}^v}$ = $\boldsymbol{\theta}\mathbf{e}^v$
$\hat{\mathbf{e}^v}$ = GradReverse.apply($\hat{\mathbf{e}^v}$, $\lambda$)
```

For instance, if the token y_t corresponds to the word "girl", we reverse the gradients for that token, and then compute additional gradients for word replacements such "child" and "kid". This is done while keeping the other gradients as they are. We update $\boldsymbol{\theta}$ by averaging the gradients of all words after the reversion. Based on our experiments, we observe that averaging all gradients stabilizes the training and yields the best results. For building a dictionary of gender words, we follow previous works (Hendricks et al., 2018; Tang et al., 2020) to use a rule-based method.

4 Experiments

In this section, we report debiasing and captioning performances on the COCO dataset and show the effectiveness of the method through qualitative results.

4.1 Implementation Details

We see that adversarial training procedure can suppress other world knowledge leading to worse generations and that further optimization improvements are necessary. We rely on two additional methods to ensure that our solution works without any degradations in the captioning performance: gender caption re-writing and identity matrix regularization.

Gender Caption Re-Writing. For each caption that contains a gender term, we follow the work (Tang et al., 2020) to replace the gender word with a corresponding gender-neutral word such as person or human, and write a new caption as additional training sample. Having neutralized captions for training is critical to our setup because it resolves training and validation discrepancy. During training, gender captions implicitly introduce dependencies between gender words and other words. During inference, the mask would discourage generating gender words and potentially affect the decoder self-attention.

Initialization & Regularization. We rely on two techniques to improve the optimization. First is initializing $\boldsymbol{\theta}$ as an identity matrix, i.e., feeding image embeddings as they are. This initializes the weights to a previous optimum, without the adversarial training. Further, we add an L1 norm penalty on the difference between $\boldsymbol{\theta}$ and the identity matrix, $\|\boldsymbol{\theta} - \mathrm{I}_K\|_1$ where $\|\cdot\|_1$ is the element-wise absolute sum, and minimize over the combined loss with the training objective in Equation 1.

4.2 Training Detail

We rely on the LAVIS package (Li et al., 2022a) to implement BLIP. We learn $\boldsymbol{\theta}$ with a batch size of 32 on eight V100 GPUs. We adopt AdamW for optimization and initialize the learning rate to be 2e-6 with linear warmup cosine annealing. We truncate captions to keep 20 words and pad them if less, and then add "A photo of" to all captions as prefix. We use the checkpoint at the fifth epoch for experiments and analysis. On the COCO caption datasets, it takes six hours to finish training for five epochs.

	BA↓	MR↑
Annotation	-0.211	0
BLIP$_{\text{ViT-L}}$	-0.239	-0.05
NeutralOut$_{\text{ViT-L}}$	-0.620	0.218
Mask	-0.619	0.207

Table 1: Results for bias amplification (BA) and gender erasing rate(ER).

4.3 Performance on Erasing Gender Bias

There are several fairness metrics used in previous works, such as Gender Ratio & Error (Hendricks et al., 2018), Bias Amplification (BA) (Zhao et al., 2017), Directional Bias Amplification (DBA) (Wang and Russakovsky, 2021), and LIC (Hirota et al., 2022b). However, some metrics are not directly applicable in this work because the proposed mask will encourage BLIP to generate degenderized captions, whereas DBA, Gender Ratio & Error, and LIC measure generated gender words. Thus, we report the BA metric, which measures the difference of gender-object correlation between training and inference. A model can amplify bias by making certain predictions at a higher rate for some groups than is to be expected based on statistics of the training data (Hall et al., 2022). We also report masking ratio (MR) on gender words, defined as the proportion of gender-related captions being de-genderized after applying the mask.

We compare the proposed mask solution with annotation and BLIP. Annotation represents the human annotated captions, which shows the difference of gender-object correlation between the training set and the validation set. BLIP$_{\text{ViT-L}}$ stands for the generated caption obtained from BLIP fine-tuned checkpoint. We consider an additional method, NeutralOut, which uses the same gender replacing rule as in Equation **??** on BLIP generated captions. Notably, we follow the work (Tang et al., 2020) when designing the rule set, so we can assume the rule set is complete and accurate. Thus, NeutralOut serves as an upper-bound for BA and MR metrics, as all gender words are replaced.

Ideally, bias amplification should be zero if the model learns the gender-object correlation well from the training set. Since the mask hides gender knowledge from the model, the gender-object

	BLEU4↑	METEOR↑	ROUGE$_L$ ↑	CIDEr↑	SPICE↑
UpDn (*)	36.6	27.7	57.5	117.0	n/a
NIC+Equalizer (*)	27.4	23.4	50.2	83.0	n/a
BLIP$_{\text{pre-train}}$ (*)	29.1	23.5	53.0	97.6	17.7
BLIP$_{\text{ViT-L}}$ (*)	**40.4**	**31.1**	**60.6**	**136.7**	**24.3**
NeutralOut$_{\text{pre-train}}$	26.6	22.9	51.3	90.5	16.7
GPTRewrite$_{\text{pre-train}}$	19.9	18.7	44.7	60.6	12.4
NeutralOut$_{\text{ViT-L}}$	37.9	**30.3**	58.8	125.9	23.1
GPTRewrite$_{\text{ViT-L}}$	32.2	26.2	53.8	99.6	18.8
Mask$_{\text{load-pre-train}}$	36.9	28.3	57.6	121.6	21.6
Mask$_{\text{load-ViT-L}}$	**38.9**	30.2	**59.2**	**131.3**	**23.2**

Table 2: Results on accuracy metrics. The higher the better. Results with (*) are taken from the respective paper.

	BLEU4↑	CIDEr↑	SPICE↑
BLIP$_{\text{ViT-L}}$	43.6	132.3	24.3
NeutralOut$_{\text{ViT-L}}$	35.3	109.6	21.4
Mask	37.4	123.0	21.3

Table 3: Results for images with human objects.

	BA↓	MR↑	BLEU4↑	CIDEr↑
NO$_{\text{ViT-L}}$	-0.620	0.218	37.9	125.8
NO$_{\text{ViT-L}}$ - 10%	-0.413	0.187	38.2	126.2
NO$_{\text{ViT-L}}$ - 20%	-0.399	0.148	38.6	128.2
NO$_{\text{ViT-L}}$ - 30%	-0.326	0.127	38.8	128.7
Mask	-0.619	0.207	38.9	131.3

Table 4: Results for simulating an incomplete rule set. NO is short for NeutralOut.

correlation would not be reflected during inference, and we would see a negative value for bias amplification. Higher masking ratio is better as a high MR means more gender words are replaced with gender-neutral words. According to Table 1, all three methods report negative bias amplification, and Mask shows the closest to NeutralOut. Since Mask is designed to hide gender information, the gender-object correlation barely exists on the generated captions. The gender masking ratio shares a similar trend as bias amplification. Both metrics clearly indicate that the proposed mask can effectively suppress gender words. Surprisingly, BLIP also slightly improve BA and MR by presenting less portion of gender words. The captioner and filter leveraged in BLIP were designed to mitigate noise web-crawled text-image pairs, which might also contribute to model debiasing.

4.4 Impact on Generated Caption

While previous experiments demonstrate the patch network can successfully mask gender knowledge, it is important to quantify if the captioning performance gets affected. We report common captioning metrics, including BLEU-4 (Papineni et al., 2002) which measures n-gram precision with a length penalty against a corpus of annotations, CIDEr (Vedantam et al., 2014) which compares cosine similarity against annotations on term frequency-inverse document frequency, and SPICE (Anderson et al., 2016) which focuses exclusively on semantic meaning, neutral translation metric METEOR (Banerjee and Lavie, 2005) which leverages wordnet synonym to compare unigram, and summarization metric ROUGE$_L$ (Lin, 2004) which measures the longest common Subsequence.

we consider the bottom-up top-down work (UpDn) (Vaswani et al., 2017) and NIC+Equalizer (Hendricks et al., 2018) as baselines. Besides NeutralOut, we further design GPTRewrite which leverages GPT model to rewrite the BLIP captions by using the prompt "There is a [BLIP CAPTION]. Rewrite it to erase gender information." (Brown et al., 2020b). The subscript means which BLIP checkpoint is being used. We report captioning metrics on the full validation set in Table 2 and a subset which includes human objects in Table 3. Based on the results we make the following observations: (i) After applying the mask to BLIP, the generated caption quality is decreased compared to BLIP, and the degradation is consistent across all metrics. This suggests that when erasing gender knowledge, the mask might hide other knowledge as well and therefore negatively affect the captioning performance. This degradation is more significant on images with human objects. (ii) The mask reports comparable results against

UpDn, indicating even though masking image representation degrades BLIP performance, it is still a strong image captioning method. Since we only introduce a simple linear layer with $K \times K$ parameters, a more complex module might bridge the performance gap, which we leave exploring other architecture in the future. (iii) The mask yields better results than NeutralOut$_{\text{ViT-L}}$ on BLEU4 and CIDEr, especially on images with human objects. NeutralOut$_{\text{ViT-L}}$ serves as the oracle for gender hiding but naively replaces words might corrupt the caption readability. Thus, although NeutralOut outperforms Mask in Table 1, one would favor Mask because it generates more natural captions. (iv) Mask also outperforms the other debiasing method, NIC+Equalizer, by a large margin, because we build the mask on top of BLIP. Given the simplicity of the introduced solution, we can apply the idea to any image representation, and expect the performance to scale with its base model.

4.5 Caption Accuracy v.s. Gender Erasing

To better understand how our Mask solution leverages the trade-off between caption accuracy and gender erasing, we manipulate the mask by combining an identity matrix and the learned $K \times K$ parameters. We introduce the hyper-parameter α and update the mask as given in Equation 2, where I_K is the identity matrix:

$$\hat{\boldsymbol{\theta}} = \alpha\boldsymbol{\theta} + (1-\alpha)\mathrm{I}_K. \tag{2}$$

We vary α from 0 to 1 and increase it by 0.1 each time. When $\alpha = 0$, the mask would be "turned-off" and $\hat{\boldsymbol{\theta}}$ would report the same result as BLIP. We plot the results on BA, MR, BLEU4, CIDEr, and SPICE w.r.t α in Figure 2. As we initialize $\boldsymbol{\theta}$ as identity matrix, Figure 2 demonstrates that Mask sacrifices caption accuracy, with a 2%-5% drop in various metrics, in return for gender erasing. Interestingly, when $\alpha = 0.25$, we see no degeneration in accuracy metrics, with an improvement in the error reduction rates. This provides a more pareto-optimal model than the baseline BLIP.

4.6 Simulating an Incomplete Rule Set

While previous experiments assume that we can find the exhaustive list of replacing rules, this would not be the case for real-world applications. For example, if we have a caption "Jenny is holding a basketball" as in (Vedantam et al., 2014), the current rule set would fail, and we need to design another rule to match "Jenny" as a female name. We simulate incomplete and inaccurate rule set by randomly dropping $K\%$ matches for the NeutralOut method, so that some gender words are not replaced. We choose $K \in [10\%, 20\%, 30\%]$ and report the results in Table 4.

According to Table 4, while NeutralOut with a lower drop rate suggests better results on BA and MR, the caption accuracy gets worse. Further, once we drop 10% matches, Mask starts to outperform NeutralOut on BA and MR and still maintains the lead on BLEU4 and CIDEr. Since Mask operates on image directly, it is more generalizable to identify male and female concepts and debias the terms corresponding to it. Thus, we conclude that Mask is more helpful when rule set is incomplete, and has potential for cases where the image has features that are more easily identifiable as male/female rather than the text.

4.7 Ablation Study

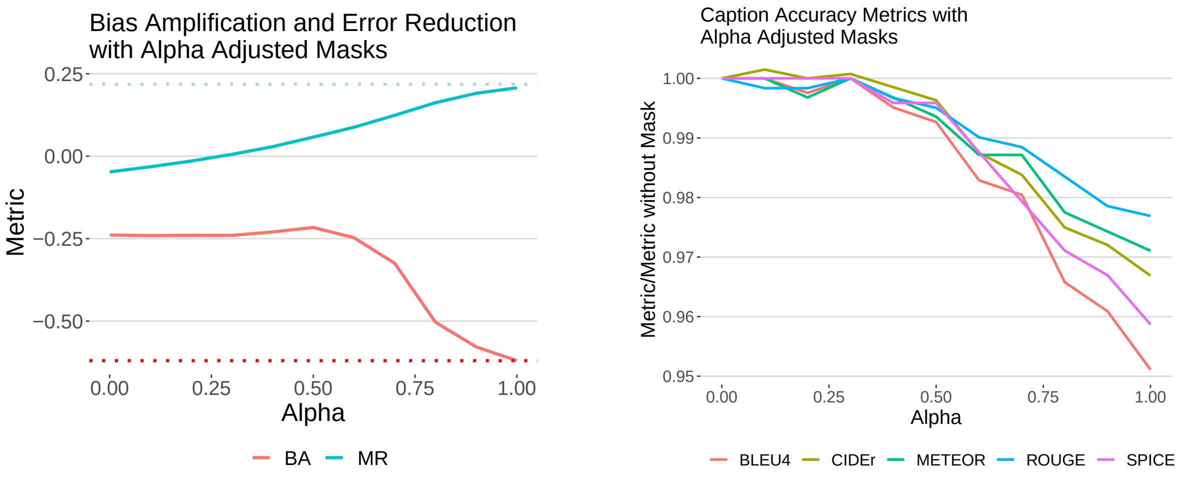
In this section, we perform an ablation study to quantify the impact of each component. We report performance on both gender bias and caption quality metrics and remove the following variants one at a time:

- w/o Neutralize Target: Implements the mask without training on de-genderized captions. This setup would introduce training and validation discrepancy as the model infers on gender words during training whereas neutral words during validation.
- w/o Negating Gradient: Implements the mask without the adversarial training. Thus, hiding gender knowledge would rely on learning from those de-genderized captions.
- w/o Identity Initialization: Randomly initializes the $K \times K$ weights instead of an identity matrix.
- w/o Identity Constraint: Does not add the L1 norm on the difference between $\boldsymbol{\theta}$ and its identity matrix.
- Large Gender Gradient: Scales up the gradient of gender words during adversarial training.

Training w/o Neutralized Targets demonstrates the importance of leverage de-genderized

	BA	MR	BLEU4	METEOR	ROUGE$_L$	CIDEr	SPICE
Mask	-0.619	0.207	38.9	30.2	59.2	131.3	23.2
w/o Neutralize Target	-0.595	0.186	27.74	23.16	48.51	94.47	18.84
w/o Negating Gradient	-0.471	0.148	37.2	28.9	58.5	126.7	22.3
w/o Identity Initialization	-0.620	0.223	23.1	20.1	44.5	74.0	14.9
w/o Identity Constraint	-0.576	0.174	36.8	28.2	57.8	121.4	21.6
Larger Gender Gradient	-0.620	0.229	15.8	15.9	37.8	48.7	10.9

Table 5: Ablation Studies: Results on debiasing metrics and accuracy metrics for variants of the mask solution.



(a) Bias metrics for the combined mask. The dotted lines represent the results of the upper bound, NeutralOut. We see that masking reduction consistently increases, although bias amplification scores are not impacted until α=0.5.

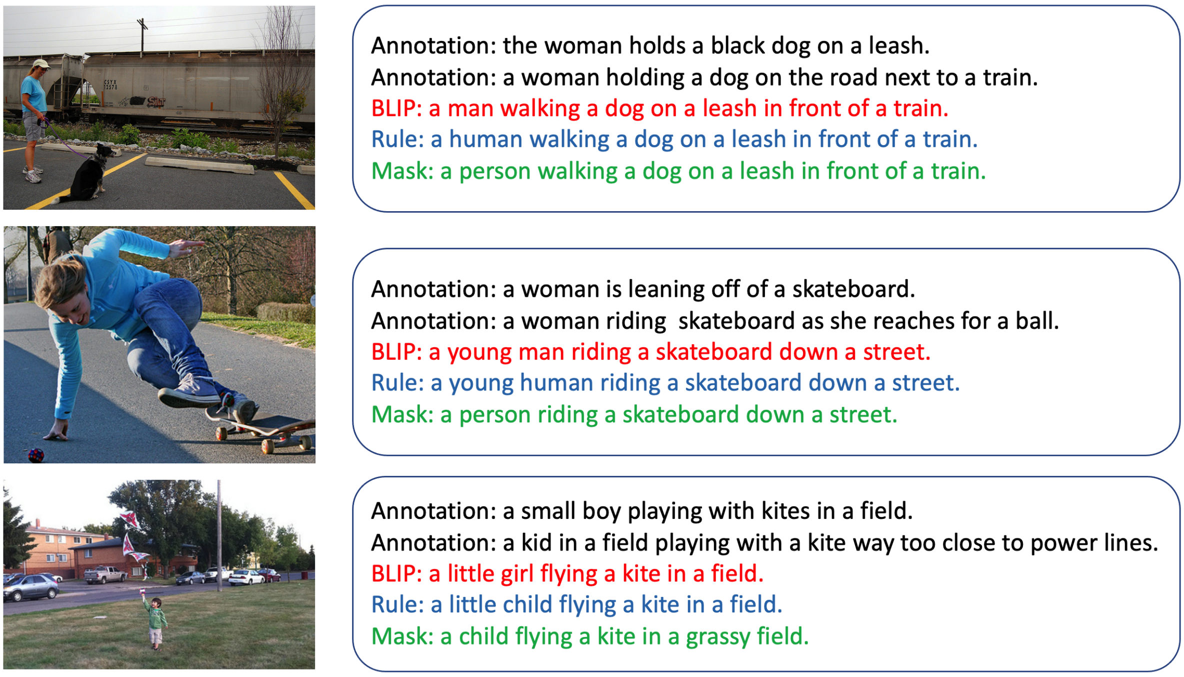
(b) Accuracy metrics for the combined mask, as presented as a percentage of the non-masked metric. We see a 2%-5% drop in various metrics at $\alpha = 1$. Further, setting $\alpha = 0.25$ results in a model that has the same performance as the baseline.

Figure 2: Bias (left) and accuracy (right) metrics for the combined mask. In implementation, the Alpha parameter can be tuned to trade off bias for accuracy depending on the use case requirements.

caption as more training examples because during training non-gender words build attention on gender words while this is not the case during inference. **Training w/o Negating Gradient** yields a strong performance on caption quality and the worst result on gender knowledge masking. Having de-genderized caption as training target serves as a fine-tuning process, and the mask can be viewed as extremely naive perceiver sampler (Alayrac et al., 2022) or adaptor (Yan et al., 2022) that have been used to align visual-text representations, which could explain the performance. Both **training w/o Identity Initialization** and using **Larger Gender Gradient** report poor captioning quality. This suggests that without suitable initialization the model would be likely to overfit on gender masking and corrupt the optimum on the image captioning task. The **Identity Constraint** seems to a large impact on all of the metrics and appears to provide significant stabilization to the optimization scheme.

4.8 Qualitative Results

We find some examples for which BLIP predicted the wrong gender class, and we present three random choices among them. We list their corresponding captions generated by the mask and compare them with BLIP caption, BLIP caption with rule-based replacement, and annotated captions in Figure 3. We see that Mask is able to generate readable captions and capture salient objects in the image. Notably, the proposed Mask could maintain the generated captions to be almost the same as the base model except the gender words, making it reliable as the rule-based method. In addition, the rule-based method could overlook non-gender words and break the readability of the caption, such as the redundant phrasing “little child" shown in Figure 3. The proposed Mask not only removes the redun-



Annotation: the woman holds a black dog on a leash.
Annotation: a woman holding a dog on the road next to a train.
BLIP: a man walking a dog on a leash in front of a train.
Rule: a human walking a dog on a leash in front of a train.
Mask: a person walking a dog on a leash in front of a train.

Annotation: a woman is leaning off of a skateboard.
Annotation: a woman riding skateboard as she reaches for a ball.
BLIP: a young man riding a skateboard down a street.
Rule: a young human riding a skateboard down a street.
Mask: a person riding a skateboard down a street.

Annotation: a small boy playing with kites in a field.
Annotation: a kid in a field playing with a kite way too close to power lines.
BLIP: a little girl flying a kite in a field.
Rule: a little child flying a kite in a field.
Mask: a child flying a kite in a grassy field.

Figure 3: Qualitative examples showing annotated and generated captions. We present three images for which BLIP has predicted the wrong gender class.

dant “little" but also adds an extra description for field. This shows potential benefit of Mask: when the caption generator is masked and does not focus on the gender terms, it focuses on other salient parts of the image and describes that in more detail.

While BLIP makes mistakes on gender, the mask solution removes gender knowledge from the image representation and prevents generation of gender-related words. Based on the three examples, Mask sacrifices caption accuracy since it cannot reveal gender information but reduces the risk of biasing one gender over the other as hypothesized.

5 Conclusion

In this work, we study the task of debiasing image captioning models. Different from existing works, we propose to mitigate gender bias by hiding gender knowledge from an image captioning model. As a result, generated captions contain gender-neutral words instead of gender words. We achieve this via applying a light-weight mask to the image embeddings.

Although we demonstrate the results on the BLIP model, the approach can be applied to any other vision-language model that ingests embeddings. As the model is frozen during training of the mask, the mask can be turned off, or tuned down (as in Section 4.5); this creates a switch with which the model owner can control the model's behavior.

Further, in order to ensure no performance degradation after debiasing, we propose an adversarial training procedure that can be generalized to other fairness/bias use cases beyond gender de-biasing.

On the COCO caption dataset, we empirically demonstrate that 1. the mask successfully masks gender knowledge; 2. our solution maintains reasonable performance on image captioning. Our analysis further suggests that it is critical to initialize the patch as an identity matrix and calibrate the training with more de-genderized captions, while further leveraging adversarial training produces the best model.

There are also a few limitations of the method. First, we observe degradation on the generated captions when comparing with BLIP. Second, we only experiment by masking with BLIP model, while theoretically the mask can be applied to any image representation. Third, we only explore the image captioning task. To address these limitations, we plan to explore other designs for the mask in future work, for instance by training masks separately for different objectives and then combining them to reduce bias across multiple cohorts. Another open question is on how well the mask idea generalizes to other solutions, and what other optimization techniques might be necessary to obtain similar performances.

6 Limitations and Ethical Considerations

This work studies gender bias in large multimodal models, specifically BLIP, on the image captioning task. The approach could degrade the overall captioning accuracy of the model by hiding not only gender but also other information from the image embedding as well. However, erasing a concept from the a model is often observed to have side effect of unlearning other information. Additional effort such as training on the degraded samples could be used to mitigate the issue.

In addition, while debiasing the gender bias, we must pay attention and not replicate the fairness issue of Gemini model. For example, we need to respect historical events and be faithful to the history. Thus, we carefully designed our experiments to adhere to ethical principles and report both debiasing metrics and utility metrics on public datasets. We compared our method against several baselines and provided thorough analysis to ensure the conclusion solid. We are presenting not just a technical improvement, but also how to reduce the risk of large models offending model users due to gender bias.

References

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. 2021. Evaluating CLIP: towards characterization of broader capabilities and downstream implications. *CoRR*, abs/2108.02818.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Christine Basta and Marta R. Costa-jussà. 2021. Impact of gender debiased word embeddings in language modeling. *CoRR*, abs/2105.00908.

Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022*, pages 806–822. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Shikha Bordia and Samuel Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. 2023. Debiasing vision-language models via biased prompts. *CoRR*, abs/2302.00070.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. 2015. Domain-adversarial training of neural networks. In *Journal of machine learning research*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. USA*, 115(16):E3635–E3644.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1012–1023. Association for Computational Linguistics.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. 2022. A systematic study of bias amplification. *CoRR*, abs/2201.11706.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, page 793–811, Berlin, Heidelberg. Springer-Verlag.

Yusuke Hirota, Yuta Nakashima, and Noa García. 2022a. Gender and racial bias in visual question answering datasets. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022b. Quantifying societal bias amplification in image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13440–13449. IEEE.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Anne Lauscher, Tobias Lüken, and Goran Glavas. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4782–4797. Association for Computational Linguistics.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. 2022a. Lavis: A library for language-vision intelligence.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022b. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5502–5515. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Tsung-Yi Lin, Michael Maire, Serge J. 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*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artif. Intell.*, 304:103654.

Thomas Manzini, Yao Chong Lim, Alan W. Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 615–621. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1878–1898. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Eric Mitchell, Peter Henderson, Christopher D. Manning, Dan Jurafsky, and Chelsea Finn. 2022a. Self-destructing models: Increasing the costs of harmful dual uses in foundation models. *CoRR*, abs/2211.14946.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022b. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Shikhar Murty, Christopher D. Manning, Scott M. Lundberg, and Marco Túlio Ribeiro. 2022. Fixing model bugs with natural language patches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11600–11613. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44.

Yao Qiang, Chengyin Li, Marco Brocanelli, and Dongxiao Zhu. 2022. Counterfactual interpolation augmentation (CIA): A unified approach to enhance fairness and explainability of DNN. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 732–739. ijcai.org.

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. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shibani Santurkar, Dimitris Tsipras, Mahalaxmi Elango, David Bau, Antonio Torralba, and Aleksander Madry. 2021. Editing a classifier by rewriting its prediction rules. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 23359–23373.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Trans. Assoc. Comput. Linguistics*, 9:1408–1424.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang.

2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, and Xia Hu. 2020. Mitigating gender bias in captioning systems. *Proceedings of the Web Conference 2021.*

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Angelina Wang, Solon Barocas, Kristen Laird, and Hanna M. Wallach. 2022. Measuring representational harms in image captioning. *2022 ACM Conference on Fairness, Accountability, and Transparency.*

Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In *International Conference on Machine Learning.*

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1995–2008. Association for Computational Linguistics.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2018. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5309–5318.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *CoRR*, abs/2010.06032.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2022. Video-text modeling with zero-shot transfer from contrastive captioners. *ArXiv*, abs/2212.04979.

Ruichen Yao, Ziteng Cui, Xiaoxiao Li, and Lin Gu. 2022. Improving fairness in image classification via sketching. *CoRR*, abs/2211.00168.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14810–14820.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 629–634. Association for Computational Linguistics.

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. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2979–2989. Association for Computational Linguistics.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI ethics of chatgpt: A diagnostic analysis. *CoRR*, abs/2301.12867.

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.