Evaluating and Mitigating Societal Bias in Image Captioning

Submitted to

Graduate School of Information Science and Technology

Osaka University

January, 2025

Yusuke HIROTA

Abstract

Societal biases in artificial intelligence, particularly in computer vision, have become a critical concern due to their potential to perpetuate and amplify harmful stereotypes. Among various computer vision tasks, image captioning exemplifies these challenges due to its interpretive nature. Image captioning models often inherit and amplify biases present in their training data, manifesting as stereotypical associations or incorrect predictions of demographic attributes such as gender and race. These biases not only compromise the fairness and reliability of generated captions but also raise ethical and societal concerns.

This thesis tackles the issue of societal bias in image captioning through three key contributions. First, it introduces a metric for measuring societal bias amplification in image captioning models. This metric quantifies how much the models amplify biases compared to their training data, providing valuable insights into the propagation of bias. Second, it proposes LIBRA, a model-agnostic framework to mitigate bias amplification in image captioning. LIBRA addresses two distinct types of gender bias—context-to-gender bias and gender-to-context bias—ensuring that efforts to reduce one do not inadvertently amplify the other. Finally, the thesis presents a dataset-level bias mitigation framework using text-guided inpainting techniques. This approach creates synthetic datasets with group-independent attribute distributions, reducing spurious correlations and enhancing fairness without sacrificing model performance.

Through these contributions, the thesis advances the understanding of societal biases in image captioning and proposes practical solutions for their quantification and mitigation. The findings pave the way for more equitable and inclusive applications of AI technologies in image captioning and beyond.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Societal biases in artificial intelligence, particularly in computer vision, have emerged as a critical challenge, posing significant risks of perpetuating and amplifying harmful stereotypes [16–21]. These biases arise when AI systems make decisions or generate outputs that systematically favor or disadvantage certain demographic groups. For example, facial recognition models trained on imbalanced datasets often misidentify individuals from underrepresented groups, leading to higher error rates for people with darker skin tones or women compared to lighter-skinned men [22]. Similarly, object classification systems have been shown to associate objects like kitchen utensils disproportionately with women, perpetuating outdated gender roles [23]. Such biases are not confined to a single task but permeate diverse applications, raising questions about the fairness and reliability of AI systems.

The manifestation of societal biases in computer vision takes many forms. In pedestrian detection, models may exhibit higher detection rates for lighter-skinned individuals compared to those with darker skin tones, affecting public safety and accessibility [24]. In autonomous vehicles, systems trained on biased datasets may misclassify or fail to recognize individuals from underrepresented demographic groups, leading to potentially life-threatening outcomes [25]. These biases often stem from imbalances in the training data, where certain groups are either underrepresented or misrepresented, as well as from model architectures that prioritize accuracy over fairness. As a result, AI systems may not perform equitably across different

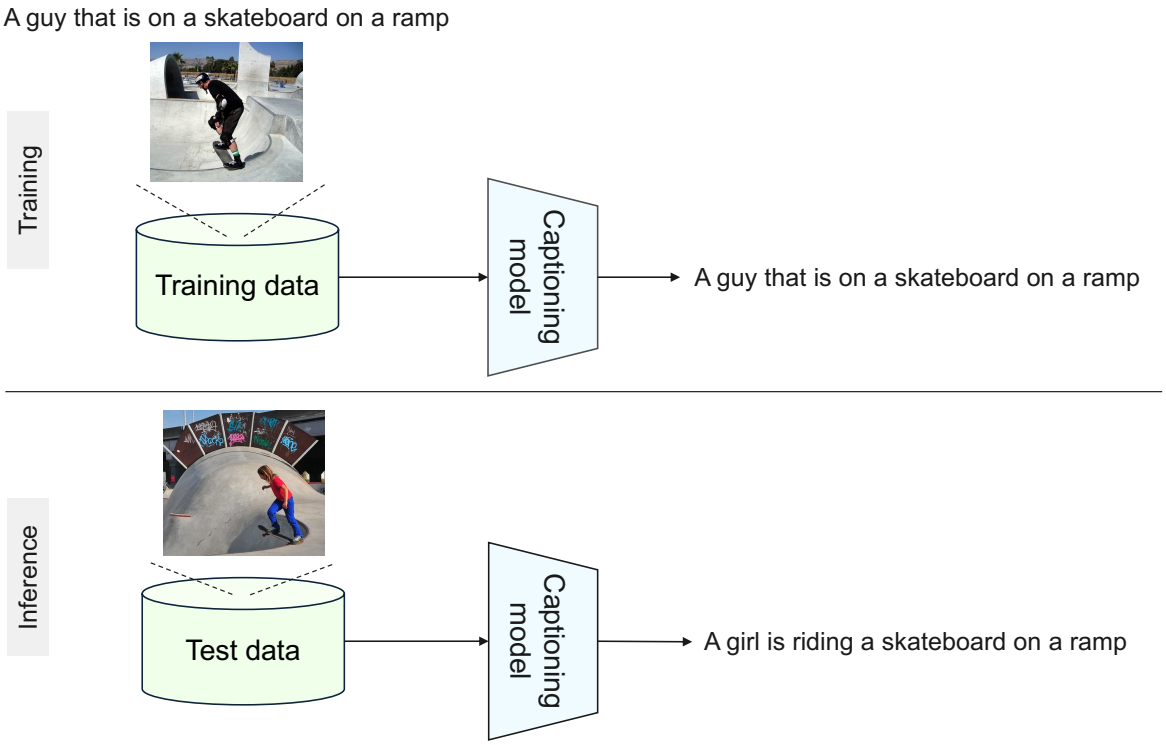


Figure 1.1: Diagram of image captioning. In the training phase, given an image and a ground-truth caption, a model is trained to predict the ground-truth captions. In the inference phase, given an image, a model generates a description about the image.

populations, reinforcing systemic inequalities.

The consequences of societal bias in computer vision extend far beyond technical errors. They can exacerbate discrimination, marginalize vulnerable groups, and erode public trust in AI technologies. For example, biased hiring algorithms have been shown to favor male candidates over equally qualified female applicants, while biased healthcare systems may misdiagnose conditions in women or people of color [26]. These issues raise profound ethical concerns, emphasizing the need for AI systems that are fair, inclusive, and equitable. Addressing these challenges is not only a technical necessity but also a societal responsibility, requiring innovative solutions to identify, measure, and mitigate biases across diverse applications.

Image captioning and societal biases Image captioning [27] is a core task in computer vision that involves generating textual descriptions for images (Figure 1.1). By bridging visual perception and natural language understanding, image captioning enables applications such as aiding visually impaired individuals, organizing digital media, and enhancing content generation. For example, given an image of a dog sitting on a sofa, the model might generate the caption "A dog relaxing on a couch." Training these models requires large-scale datasets, such as MSCOCO [28], where each image is paired with human-annotated captions. Despite achieving state-of-the-art accuracy, these systems are not immune to biases inherent in their training data, which can significantly impact their fairness and reliability [3].

Unlike other computer vision tasks, such as object detection or classification, image captioning faces unique challenges regarding societal bias due to its interpretive nature. These biases manifest in various forms. For instance, models often stereotype women as being associated with domestic activities, generating captions like "A woman in a kitchen" even when the image actually shows a man in the kitchen (Figure 1.2 (top right)). This occurs because the training data strongly correlates kitchen-related objects with women, leading the model to incorrectly infer gender based on the surrounding context rather than the actual individual in the image [3]. Similarly, models may produce captions such as "A man wearing a suit" even when the man in the image is not wearing a suit (Figure 1.2 (bottom left)). This happens because the training data frequently associates men with suits in professional settings, causing the model to overgeneralize and generate captions that reinforce these stereotypes [6, 29]. These biases are not just reproduced from the training data but are often amplified—a phenomenon known as bias amplification.

The causes of societal bias in image captioning stem from both data and model-related factors. At the dataset level, imbalances occur when certain groups are overrepresented or underrepresented in specific contexts. For example, datasets may contain more images of men in sports-related activities and women in domestic ones, causing the model to learn skewed associations [6]. At the model level, the reliance on statistical correlations to infer contextual meanings can lead to overgeneralizations [6, 29]. For instance, a model trained to associate

the presence of kitchen objects with women might predict "A woman standing in the kitchen" regardless of whether a person is visible in the image. These biases not only compromise fairness but also degrade the descriptive accuracy and reliability of the generated captions.

Addressing these biases requires systematic approaches to both quantify and mitigate their effects. Quantification involves developing robust metrics to measure bias and its amplification in image captioning outputs. These metrics help to understand how much the models amplify societal biases, providing a foundation for targeted interventions. Mitigation strategies, on the other hand, focus on reducing bias during model training and inference. This thesis contributes to this effort by proposing novel methods for both quantification and mitigation, aiming to create fairer and more inclusive image captioning systems. By tackling these challenges, this thesis advances the understanding of societal biases in AI and paves the way for more equitable applications of image captioning technology.

Understanding and quantifying bias in image captioning models Bias in image captioning models is not limited to reproducing inequalities in training datasets; it often goes further by amplifying these biases during generation. For instance, as shown in the first study of this thesis, even state-of-the-art image captioning models trained on large-scale datasets like MSCOCO exhibit substantial gender and racial biases. These biases manifest in skewed word associations and stereotypical descriptions that reflect and reinforce societal inequalities. Moreover, current evaluation metrics often fail to fully capture the extent of these biases, highlighting the need for comprehensive and unified approaches to bias quantification. This work introduces a novel bias measurement metric that evaluates the amplification of societal biases in image captioning models.

Mitigating bias amplification in image captioning models Building upon the evaluation of bias amplification, the second study focuses on designing a framework to mitigate biases in image captioning models. The proposed method, LIBRA, addresses two key types of gender bias: context-to-gender context $\rightarrow$ gender bias and gender-to-context gender $\rightarrow$ context bias.

LIBRA synthesizes biased captions to expose models to biased scenarios during training and trains a debiasing caption generator to recover unbiased captions. Unlike prior methods, LIBRA considers both types of biases simultaneously, ensuring that efforts to reduce one bias do not inadvertently amplify another. Experimental results demonstrate LIBRA's effectiveness in mitigating gender bias across multiple metrics, offering a robust, model-agnostic solution to improve fairness in image captioning systems.

Towards dataset-level bias mitigation While model-level solutions like LIBRA play a crucial role, addressing bias at the dataset level is equally important. The third study in this thesis introduces a novel framework for generating synthetic datasets with group-independent attribute distributions using text-guided inpainting techniques. This approach reduces spurious correlations between protected attributes (e.g., gender, skin tone) and visual content, ensuring fairer training conditions. By decorrelating both labeled and unlabeled attributes, this method significantly reduces biases in image captioning and classification tasks without sacrificing model performance. Importantly, the framework addresses ethical concerns by modifying only masked image regions and employs rigorous data filtering to ensure the quality and fidelity of synthetic data.

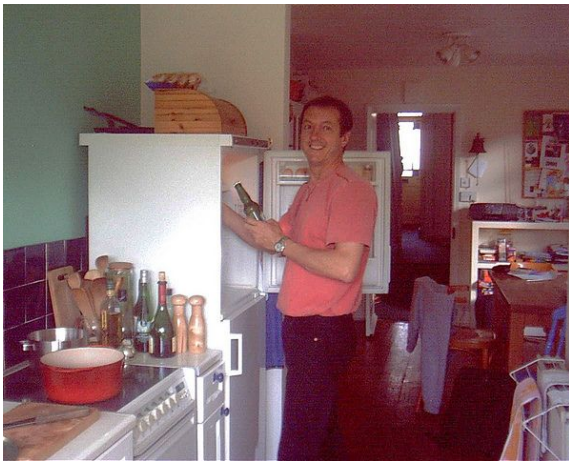
Contributions and Roadmap This thesis makes the following key contributions to the field of bias mitigation in image captioning:

1. A novel metric for quantifying societal bias amplification in image captioning models.
2. A model-agnostic debiasing framework, LIBRA, that effectively mitigates multiple types of bias in caption generation.
3. A synthetic dataset generation pipeline that decorrelates attributes from protected groups, addressing biases at the dataset level.

The remainder of this thesis is organized as follows: Chapter 2 presents the bias measurement metric and its applications. Chapter 3 details the LIBRA method and its experimental

validation. Chapter 4 introduces the dataset-level bias mitigation approach, including its implementation and evaluation. Chapter 5 discusses the relationship between computer vision research and social science perspectives on societal bias.

Baseline
a **man** sitting on a motorcycle



Baseline
a **woman** standing in a kitchen

Baseline
a man wearing a **suit** and a tie

Baseline
a woman in a colorful **dress**

Figure 1.2: Examples of gender bias in image captioning models. Captioning models can predict incorrect gender words based on the stereotypical context for the gender (top row). Additionally, models can generate gender-stereotypical words (bottom row).

Chapter 2

Quantifying Societal Bias Amplification in Image Captioning

2.1 Overview

The presence of undesirable biases in computer vision applications is of increasing concern. The evidence shows that large-scale datasets, and the models trained on them, present major imbalances in how different subgroups of the population are represented [3, 22, 23, 30]. Detecting and addressing these biases, often known as societal biases, has become an active research direction in our community [31–37].

Contrary to popular belief, the presence of bias in datasets is not the only cause of unfairness [38]. Model choices and how the systems are trained also have a large impact on the perpetuation of societal bias. This is supported by evidence: 1) models are not only reproducing the inequalities of the datasets but amplifying them [23], and 2) even when trained on balanced datasets, models may still be biased [39] as the depth of historical discrimination is more profound than what it can be manually annotated, i.e., bias is not always evident to the human annotator eye.

The prevalence of accuracy as the single metric to optimize in most popular benchmarks [40]

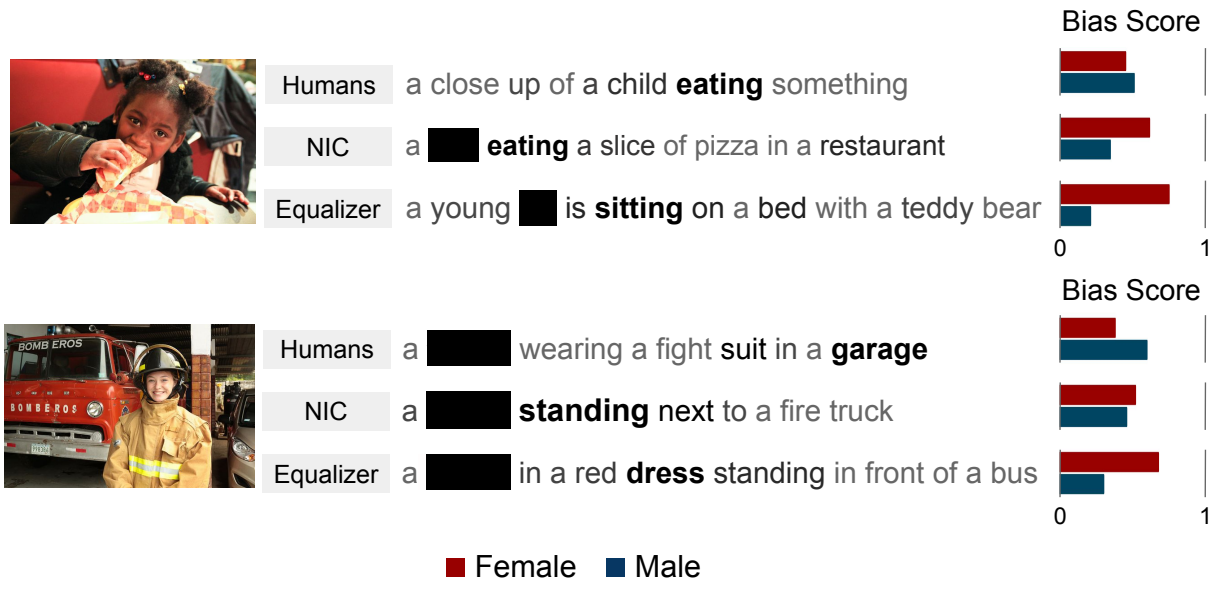


Figure 2.1: Measuring gender bias in MSCOCO captions [1]. For each caption generated by humans, NIC [2], or NIC+Equalizer [3], we show our proposed bias score for *female* and *male* attributes. This bias score indicates how much a caption is biased toward a certain protected attribute. The contribution of each word to the bias score is shown in gray-scale (bold for the word with the highest contribution). Gender revealing words are masked.

has made other aspects of the models, such as fairness, cost, or efficiency, not a priority (and thus, something to not look into). But societal bias is a transversal problem that affects a variety of tasks within computer vision, such as *facial recognition*, with black women having higher error rates than white men [22]; *object classification*, with kitchen objects being associated with women with higher probabilities than with men [23]; or *pedestrian detection*, with lighter skin individuals showing higher detection rates than darker skin people [24]. Although the causes of societal bias in different computer vision systems may be similar, the consequences are particular and require specific solutions for each task.

We examine and quantify societal bias in image captioning (Figure 2.1). Image captioning has achieved state-of-the-art accuracy on MSCOCO captions dataset [1] by means of pre-trained visual and language Transformers [5]. By leveraging very large-scale collections of data (e.g.,

Google Conceptual Captions [41] with about 3.3 million image-caption pairs crawled from the Internet), self-attention-based models [42] have the potential to learn world representations according to the training distribution. However, these large amounts of data, often without (or with minimal) curation, conceal multiple problems, including the normalization of abuse or the encoding of discrimination [30, 43, 44]. So, once image captioning models have achieved outstanding performance on evaluation benchmarks, a question arises: are these models safe and fair to everyone?

We are not the first to formulate this question. Image captioning has been shown to reproduce gender [3] and racial [6] bias. By demonstrating the existence of societal bias in image captioning, the pioneering work in [3] set the indispensable seed to continue to investigate this problem, which we believe is far from being solved. We argue that one of the aspects that remains open is the quantification and evaluation of societal bias in image captioning. So far, a variety of metrics have been applied to assess different aspects of societal bias in human and model-generated captions, such as whether the representation of different subgroups is balanced [3, 6] or whether the protected attributes[1] values (e.g., *female*, *male*) are correctly predicted [3, 45]. However, in Section 2.3, we show that current metrics may be insufficient, as they only consider the effects of bias perpetuation to a degree.

With the aim to identify and correct bias in image captioning, in Section 2.4, we propose a simple but effective metric that measures not only how much biased a trained captioning model is, but also how much bias is introduced by the model with respect to the training dataset. This simple metric allows us to conduct a comprehensive analysis of image captioning models in terms of gender and racial bias (Section 2.5), with an unexpected revelation: the gender equalizer designed to reduce gender bias in [3] is actually amplifying gender bias when considering the semantics of the whole caption. This discovery highlights, even more, the necessity of a standard, unified metric to measure bias and bias amplification in image captioning, as the efforts to address societal inequalities will be ineffective without a tool to quantify how much

[1] *Protected attribute* refers to a demographic variable (age, gender, race, etc.) that a model should not use to produce an output.

bias a system exhibits and where this bias is coming from. We conclude with an analysis of the limitation of the proposed metric in Section 2.6 and a summary of the main findings in Section 2.7.

2.2 Related work

Societal bias in computer vision The problem of bias in large-scale computer vision datasets was first raised by Torralba and Efros in [40], where the differences in the image domain between datasets were explored. Each dataset presented different versions of the same object (e.g., cars in Caltech [46] tended to appear sidewise, whereas in ImageNet [47] were predominantly of racing type), impacting cross-dataset generalization. But it was only recently that societal bias in computer vision was formally investigated.

In the seminal work of Buolamwini and Gebru [22], commercial face recognition applications were examined across subgroups, demonstrating that performance was different according to the gender and race of each individual, especially misclassifying women with darker skin tones. Similarly, Zhao et al. [23] showed not only that images in MSCOCO [28] were biased towards a certain gender, but also that action and object recognition models amplified such bias in their predictions. With an increased interest in fairness, multiple methods for mitigating the effects of dataset bias have been proposed [23, 37, 39, 48, 49].

Measuring societal bias Societal bias is a problem with multiple layers of complexity. Even on balanced datasets, models still perpetuate bias [39], indicating that social stereotypes are occurring at the deepest levels of the image. This makes the manual identification and annotation of biases unfeasible. Thus, the first step towards fighting and mitigating bias is to quantify the problem.

Bias quantification metrics have been introduced for image classification. Zhao et al. [23] defined bias based on the co-occurrence of objects and protected attributes; Wang et al. [39] relied on the accuracy of a classifier when predicting the protected attributed; and Wang and

Russakovsky [50] extended the definition of bias by including directionality. In addition, REVISE [31] and CIFAR-10S [12] ease the task of identifying bias on datasets and models, respectively. These solutions, however, cannot be directly applied to image captioning, so specific methods must be developed.

Societal bias in image captioning In image captioning [2, 4, 27, 51] the input to the model is an image and the output is a natural language sentence. This duality of data modalities makes identifying bias particularly challenging, as it can be encoded in the image and/or in the language. The original work by Burns et al. [3] showed that captions in MSCOCO [1] present gender imbalance and proposed an equalizer to force captioning models to produce gender words based on visual evidence. Recently, Zhao et al. [6] studied racial bias from several perspectives, including visual representation, sentiment analysis, and semantics.

Each of these studies, however, uses different evaluation protocols and definitions of bias, lacking of a standard metric. To fill this gap, we propose an evaluation metric to measure not only how biased a model is, but how much it is amplified with respect to the original (biased) dataset.

2.3 Analysis of fairness metrics

Bias in image captioning has been estimated using different methods: how balanced the prediction of the protected attributed is [3], the overlap of attention maps with segmentation annotations [3], or the difference in accuracy between the different protected attributes [6]. In this section, we thoroughly examine existing fairness evaluation metrics and their shortcomings when applied to image captioning.

Notation Let $\mathcal{D}$ denote the training split of a certain vision dataset with samples (I, y, a), where I is an image, y is the ground-truth annotation for a certain task, and $a \in \mathcal{A}$ is a protected attribute in set $\mathcal{A}$. The validation/test split is denoted by $\mathcal{D}'$. We assume there is a model M

that makes prediction $\hat{y}$ associated with this task from the image, i.e., $\hat{y} = M(I)$. For image captioning, we define a ground-truth caption $y = (y_1, y_2, \ldots, y_n)$ as a sequence of n tokens.

2.3.1 Fairness metrics in image captioning

Difference in performance A natural strategy to show bias in image captioning is as the difference in performance between the subgroups of a protected attribute, in terms of accuracy [3, 6, 45], ratio [3], or sentiment analysis [6]. Quantifying the existence of different behavior according to demographic groups is essential to demonstrate the existence of bias in a model, but it is insufficient for a deeper analysis, as it does not provide information on where the bias comes from, and whether bias is being amplified by the model. Thus, it is good practice to accompany difference in performance with other fairness metrics.

Attribute misclassification Another common metric is to check if the protected attribute has been correctly predicted in the generated caption [3, 45]. This assumes that the attribute can be clearly identified in a sentence, which may be the case for some attributes, e.g., age (*a young person*, *a child*) or gender (*a woman*, *a man*), but not for others, e.g., skin tone. This is critical for two reasons: 1) even when the attribute is not clearly mentioned in a caption, bias can occur through the use of different language to describe different demographic groups; and 2) it only considers the prediction of the protected attribute, ignoring the rest of the sentence which may also exhibit bias.

Right for the right reasons Introduced in [3], it measures whether the attention activation maps when generating a protected attribute word w in the caption, e.g., *woman* or *man*, are located in the image region where the evidence about the protected attribute is found, i.e., the person. This metric quantifies the important task of whether w is generated based on the person visual evidence or, on the contrary, on the visual context, which has been shown to be one of the sources of bias in image captioning models. However, it has three shortcomings: 1) it needs a shortlist of protected attribute words, and a person segmentation map per image, which may

not always be available; 2) it assumes that visual explanations can be generated from the model, which may not always be the case; and 3) it does not consider the potential bias in the rest of the sentence, which (as we show in Section 2.5) is another critical source of bias.

Sentence classification Lastly, Zhao et al. [6] introduced the use of sentence classifiers for analyzing racial bias. The reasoning is that if a classifier can distinguish between subgroups in the captions, the captions contain bias. Formally, let f denote a classifier that predicts a protected attribute in $\mathcal{A}$ trained over $\mathcal{D}$, i.e., $\hat{a} = f(y)$, from a caption y in an arbitrary set $\mathcal{H}$ of captions. If the accuracy is higher than the chance rate, y is considered to be biased:

$$\text{SC} = \frac{1}{|\mathcal{H}|} \sum_{y \in \mathcal{H}} \mathbb{1}[f(y) = a], \tag{2.1}$$

where $\mathbb{1}[\cdot]$ is a indicator function that gives 1 when the statement provided as the argument holds true and 0 otherwise. $\mathcal{H}$ typically is the set of all captions generated from the images in the test/validation split $\mathcal{D}'$ of the dataset, i.e., $\mathcal{H} = \{M(I) \mid I \in \mathcal{D}'\}$.

Unlike the previous methods, this metric considers the full context of the caption. However, a major shortcoming is that, when bias exists on the generated data, the contributing source is not identified. Whether the bias comes from the model or from the training data and whether bias is being amplified or not, cannot be concluded.

2.3.2 Bias amplification metrics

There is a family of metrics designed to measure bias amplification on visual recognition tasks. We describe them and analyze the challenges when applied to captioning.

Bias amplification Proposed in [23], it quantifies the implicit correlations between model prediction $\hat{y} = M(I)$ and the protected attribute $a \in \mathcal{A}$ by means of co-occurrence, and whether these correlations are more prominent in the model predictions or in the training data. Let $\mathcal{L}$ denote the set of possible annotations l in the given task, i.e., y and $\hat{y}$ are in $\mathcal{L}$; c_a and $\hat{c}_a$ denote

the numbers of co-occurrences of a and l, counted over y and $\hat{y}$, respectively. Bias is

$$\tilde{b}_{al} = \frac{\tilde{c}_{al}}{\sum_{a \in \mathcal{A}} \tilde{c}_{al}}, \tag{2.2}$$

where $\tilde{c}$ is either c or $\hat{c}$, and $\tilde{b}$ is either b or $\hat{b}$, respectively. Then, bias amplification is defined by

$$\text{BA} = \frac{1}{|\mathcal{L}|} \sum_{a \in \mathcal{A}, l \in \mathcal{L}} (\hat{b}_{al} - b_{al}) \times \mathbb{1}\left[b_{al} > \frac{1}{|\mathcal{A}|} \right]. \tag{2.3}$$

BA > 0 means that bias is amplified by the model, and otherwise mitigated. This metric is useful for a classification task, such as action or image recognition, for which the co-occurrence can be easily counted. However, one of the major shortcomings is that it ignores that protected attributes may be imbalanced in the dataset, e.g., in MSCOCO images [28] there are 2.25 more men than women, which causes most of objects to be correlated with men. To solve this and other issues, Wang and Russakovsky [50] proposed an extension called directional bias amplification.

Leakage Another way to quantify bias amplification is leakage [39], which relies on the existence of a classifier to predict the protected attribute a. For a sample (I, y, a) in $\mathcal{D}$ with a ground-truth annotation y, a classifier f predicts the attribute $a \in \mathcal{A}$ from either y or $\hat{y} = M(I)$. Using this, the leakage can be formally defined as,

$$\text{Leakage} = \lambda_M - \lambda_D, \tag{2.4}$$

where

$$\lambda_D = \frac{1}{|\mathcal{D}|} \sum_{(y,a) \in \mathcal{D}} \mathbb{1}[f(y) = a] \tag{2.5}$$

$$\lambda_M = \frac{1}{|\mathcal{D}|} \sum_{(I,a) \in \mathcal{D}} \mathbb{1}[f(\hat{y}) = a] \tag{2.6}$$

A positive leakage indicates that M amplifies the bias with respect to the training data, and mitigates it otherwise.

Challenges The direct application of the above metrics to image captioning presents two major challenges. Let us first assume that, for image captioning, the set of words in the vocabulary corresponds to the set $\mathcal{L}$ of annotations in Eq. (2.3) under a multi-label setting. The first challenge is that these metrics do not consider the semantics of the words: e.g., in the sentences *a woman is cooking* and *a woman is making dinner*, the tokens *cooking* and *making dinner* would be considered as different annotation l. The second challenge is they do not consider the context of each word/task: e.g., the token *cooking* will be seen as the same task in the sentence *a man is cooking* and in *a man is not cooking*.

2.4 Bias amplification for image captioning

We propose a metric to specifically measure bias amplification in image captioning models, borrowing some ideas from sentence classification [6] and leakage [39]. Our metric, named LIC, is built on top of the following hypothesis:

> ***Hypothesis 1.*** In an unbiased set of captions, there should not exist differences between how demographic groups are represented.

Caption masking As discussed in Section 2.3, for some protected attributes (e.g., age and gender), specific vocabulary may be explicitly used in the captions. For example, consider *gender* as a binary[2] protected attribute a with possible values $\{female, male\}$. The sentence

A girl is playing piano,

directly reveals the protected attribute value of the caption, i.e., *female*. To avoid explicit mentions to the protected attribute value, we preprocess captions by masking the words related to that attribute.[3] The original sentence is then transformed to the masked sentence

[2]Due to the availability of annotations in previous work, we use the binary simplification of gender, acknowledging that it is not inclusive and should be addressed in future work.

[3]A list of attribute-related words is needed for each protected attribute.

A [redacted] is playing piano.

Note that this step is not always necessary, as some protected attributes are not explicitly revealed in the captions.

Caption classification We rely on a sentence classifier f_s to estimate societal bias in captions. Specifically, we encode each masked caption y' [4] with a natural language encoder E to obtain a sentence embedding e, as $e = E(y')$. Then, we input e into the sentence classifier f_s, whose aim is to predict the protected attribute a from y' as

$$\hat{a} = f_s(E(y')) \tag{2.7}$$

E and f_s are learned on a training split $\mathcal{D}$. According to *Hypothesis 1*, in an unbiased dataset, the classifier f_s should not find enough clues in y' to predict the correct attribute a. Thus, $\mathcal{D}$ is considered to be biased if the empirical probability $p(\hat{a} = a)$ over $\mathcal{D}$ is greater than the chance rate.

Bias amplification Bias amplification is defined as the bias introduced by a model with respect to the existing bias in the training set. To measure bias amplification, we quantify the difference between the bias in the generated captions set $\hat{\mathcal{D}} = \{\hat{y} = M(I) \mid I \in \mathcal{D}\}$ with respect to the bias in the original captions in the training split $\mathcal{D}$.

One concern with this definition, particular to image captioning, is the difference in the vocabularies used in the annotations and in the predictions, due to 1) the human generated captions typically come with a richer vocabulary, 2) a model's vocabulary is rather limited, and 3) the vocabulary itself can be biased. Thus, naively applying Eq. (2.4) to image captioning can underestimate bias amplification. To mitigate this problem, we introduce noise into the original human captions until the vocabularies in the two sets (model generated and human generated) are aligned. Formally, let $\mathcal{V}_{\text{ann}}$ and $\mathcal{V}_{\text{pre}}$ denote the vocabularies identified for all annotations and predictions in the training set, respectively. For the annotation $y = (y_1, \ldots, y_N)$, where y_n

[4] If caption masking is not applied, $y' = y$.

is the n-th word in y, we replace all y_n in $\mathcal{V}_{\text{ann}}$ but not in $\mathcal{V}_{\text{pre}}$ with a special out-of-vocabulary token to obtain perturbed annotations $y^\star$, and we train the classifier $f_s^\star$ over $\{y^\star\}$.

The LIC metric The confidence score $s_a^\star$ is an intermediate result of classifier $f_s^\star$, i.e.,

$$\hat{a} = f_s^\star(y^\star) = \text{argmax}_a s_a^\star(y^\star), \tag{2.8}$$

and it can be interpreted as a posterior probability $p(\hat{a} = a \mid y^\star)$ of the protected attribute a and can give an extra hint on how much $y^\star$ is biased toward a. In order words, not only the successful prediction rate is important to determine the bias, but also how confident the predictions are. The same applies to $\hat{s}_a$ and $\hat{f}_s$ trained with $\{\hat{y}\}$. We incorporate this information into the Leakage for Image Captioning metric (LIC), through

$$\text{LIC}_D = \frac{1}{|\mathcal{D}|} \sum_{(y^\star, a) \in \mathcal{D}} s_a^\star(y^\star) \mathbb{1}[f^\star(y^\star) = a] \tag{2.9}$$

$$\text{LIC}_M = \frac{1}{|\hat{\mathcal{D}}|} \sum_{(\hat{y}, a) \in \hat{\mathcal{D}}} \hat{s}_a(\hat{y}) \mathbb{1}[\hat{f}(\hat{y}) = a], \tag{2.10}$$

so that LIC is finally computed as

$$\text{LIC} = \text{LIC}_M - \text{LIC}_D. \tag{2.11}$$

where a model is considered to amplify bias if $\text{LIC} > 0$. We refer to $\hat{s}_a$ as the *bias score*.

2.5 Experiments

Data Experiments are conducted on a subset of the MSCOCO captions dataset [1]. Specifically, we use the images with binary gender and race annotations from [6]: *female* and *male* for gender, *darker* and *lighter* skin tone for race.[5] Annotations are available for images in the validation set with person instances, with a total of $10,780$ images for gender and $10,969$ for

[5] Similarly, due to the availability of annotations in previous work, we use a binary simplification for race and skin tone. We acknowledge that these attributes are much more complex in reality.

race. To train the classifiers, we use a balanced split with equal number of images per protected attribute value, resulting in $5,966$ for training and 662 for test in gender, and $1,972$ for training and 220 for test in race. Other metrics are reported on the MSCOCO val set.

Metrics We report bias using LIC, together with LIC_D in Eq. (2.9) and LIC_M in Eq. (2.10). For gender bias, we also use Ratio [3], Error [3], Bias Amplification (BA) [23], and Directional Bias Amplification [50]. Directional bias amplification is computed for object $\rightarrow$ gender direction (DBA_G) and for gender $\rightarrow$ object direction (DBA_O) using MSCOCO objects [28]. For skin tone, we only use LIC, LIC_D, and LIC_M, as there are no words we can directly associated with race in the captions to calculate the other metrics. Accuracy is reported in terms of standard metrics BLEU-4 [8], CIDEr [52], METEOR [10], and ROUGE-L [53].

Models We study bias on captions generated by the following models: NIC [2], SAT [27], FC [54], Att2in [54], UpDn [4], Transformer [42], OSCAR [5], NIC+ [3], and NIC+Equalizer [3]. NIC, SAT, FC, Att2in, and UpDn are classical CNN [55] encoder-LSTM [56] decoder models. Transformer and OSCAR are Transformer-based [42] models, which are the current state-of-the-art in image captioning. NIC+ is a re-implementation of NIC in [3] trained on the whole MSCOCO and additionally trained on MSCOCO-Bias set consisting of images of male/female. NIC+Equalizer is NIC+ with a gender bias mitigation loss, that forces the model to predict gender words only based on the region of the person. Note that most of the pre-trained captioning models provided by the authors are trained on the Karpathy split [57], which uses both train and validation splits for training. As the val set is part of our evaluation, we retrain all the models on the MSCOCO train split only.

LIC metric details For masking, we replace pre-defined gender-related words[6] with a special token `<gender>`. We do not mask any words for race prediction because race is not commonly explicitly mentioned in the captions.

[6]The list of gender-related words can be found in the appendix.

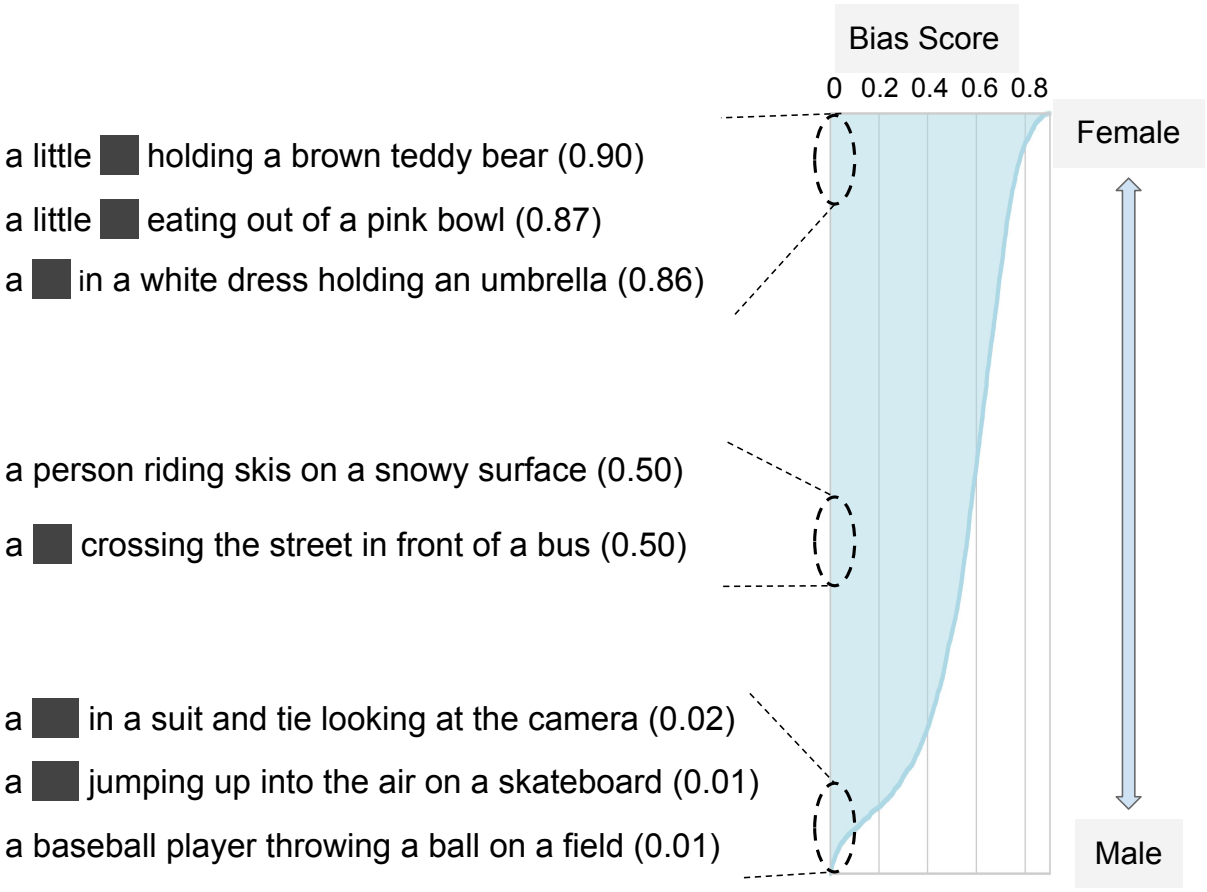


Figure 2.2: Gender bias score for captions generated with OSCAR. Masked captions are encoded with a LSTM and fed into a gender classifier. Bias score correlates with typical gender stereotypes.

The LIC classifier is based on several fully-connected layers on top of a natural language encoder. For the encoder, we use a LSTM [56] for our main results. We do not initialize the LSTM with pre-computed word embeddings, as they contain bias [58, 59]. For completeness, we also report LIC when using BERT [60], although it has also been shown to exhibit bias [61, 62] and it can affect our metric. BERT is fine-tuned (BERT-ft) or used as is (BERT-pre). The classifier is trained 10 times with random initializations, and the results are reported by the average and standard deviation.

Table 2.1: Gender bias and accuracy for several image captioning models. Red/green denotes the worst/best score for each metric. For bias, lower is better. For accuracy, higher is better. BA, DBA_G, and DBA_O are scaled by 100. Unbiased model is $\text{LIC}_M = 25$ and $\text{LIC} = 0$.

	Gender bias ↓							Accuracy ↑			
Model	LIC	LIC_M	Ratio	Error	BA	DBA_G	DBA_O	BLEU-4	CIDEr	METEOR	ROUGE-L
NIC [2]	**3.7**	**43.2**	2.47	**14.3**	4.25	3.05	**0.09**	**21.3**	**64.8**	**20.7**	**46.6**
SAT [27]	5.1	44.4	2.06	7.3	1.14	3.53	0.15	32.6	98.3	25.8	54.1
FC [54]	8.6	46.4	2.07	10.1	4.01	**3.85**	0.28	30.5	98.0	24.7	53.5
Att2in [54]	7.6	45.9	2.06	4.1	**0.32**	3.60	**0.29**	33.2	105.0	26.1	55.6
UpDn [4]	9.0	48.0	2.15	3.7	2.78	3.61	0.28	36.5	117.0	27.7	57.5
Transformer [42]	8.7	48.4	2.18	3.6	1.22	3.25	0.12	32.3	105.3	27.0	55.1
OSCAR [5]	9.2	48.5	2.06	**1.4**	1.52	3.18	0.19	**40.4**	**134.0**	**29.5**	**59.5**
NIC+ [3]	7.2	46.7	**2.89**	12.9	**6.07**	**2.08**	0.17	27.4	84.4	23.6	50.3
NIC+Equalizer [3]	**11.8**	**51.3**	**1.91**	7.7	5.08	3.05	0.20	27.4	83.0	23.4	50.2

2.5.1 LIC analysis

We qualitatively analyze the LIC metric to verify whether it is consistent with human intuition. We generate captions in the test set with OSCAR, mask the gender-related words, and encode the masked captions with a LSTM classifier to compute LIC bias score, $\hat{s}_a$, for the gender attribute, as formulated in Section 2.4. Then, we manually inspect the captions and their associated bias score.

Figure 2.2 shows generated captions with higher, middle, and lower bias scores. The bias score assigned to each caption matches typical gender stereotypes. For example, the third caption from the top, "a [redacted] in a white dress holding an umbrella", yields a very high bias score for *female*, probably due to the stereotype that the people who wear dresses and holds umbrellas tend to be women. On the contrary, the bottom caption, "a baseball player throwing a ball on a field", with one of the lowest scores assigned to *female*, perpetuates the stereotype that baseball players are mostly men. Additionally, when inspecting the captions with a bias score around 0.5, we see that the descriptions tend to be more neutral and without strong gender stereotypes. This support the importance of including $s_a^\star$ and $\hat{s}_a$ in the LIC computation, as in Eqs. (2.9) and

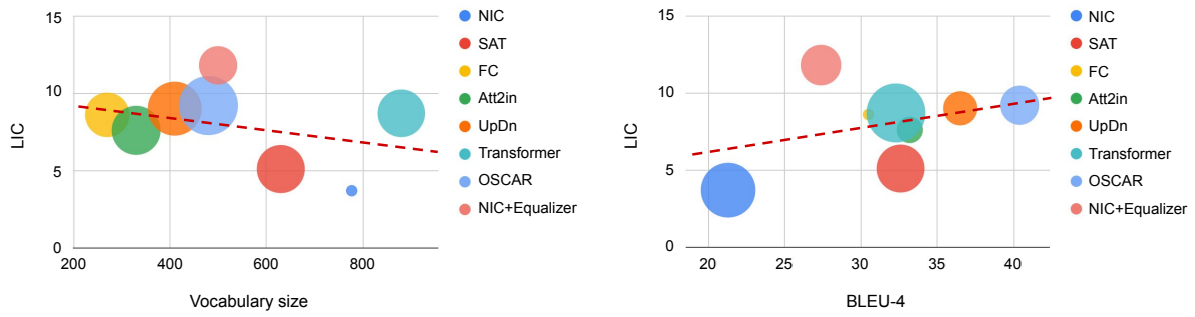


Figure 2.3: LIC vs. Vocabulary size (left) and BLEU-4 score (right). The size of each bubble indicates the BLEU-4 score (left) or the vocabulary size (right). Score tends to decrease with largest vocabularies, but increase with more accurate BLEU-4 models, whereas NIC+Equalizer [3] is presented as an outlier. The dotted lines indicate the tendency, $R^2 = 0.153$ (left) and $R^2 = 0.156$ (right).

(2.10).

2.5.2 Quantification of gender bias

We evaluate the gender bias of different captioning models in terms of LIC together with adaptations of existing bias metrics. For BA, we use the top $1,000$ common words in the captions as $\mathcal{L}$, whereas for DBA_G and DBA_O, we use MSCOCO objects [28]. More details can be found in the appendix. Results are shown in Table 2.1. We also show the relationship between the quality of a caption, in terms of vocabulary and BLEU-4 score, with LIC in Figure 2.3. Finally, we compare LIC when using different language encoders in Table 2.2. The main observations are summarized below.

Observation 1.1. **All the models amplify gender bias.** In Table 2.1, all the models have a LIC_M score well over the unbiased model ($\text{LIC}_M = 25$), with the lowest score being 43.2 for NIC. When looking at LIC, which indicates how much bias is introduced by the model with respect to the human captions, also all the models exhibit bias amplification, again with NIC having the lowest score. NIC is also the model that performs the worst in terms of accu-

Table 2.2: Gender bias scores according to LIC, LIC_M, and LIC_D for several image captioning models. Captions are encoder with LSTM, BERT-ft, or BERT-pre. Unbiased model is $\text{LIC}_M = 25$ and $\text{LIC} = 0$. It shows that LIC is consistent across different language models.

	LSTM			BERT-ft			BERT-pre		
Model	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC
NIC [2]	43.2 ± 1.5	39.5 ± 0.9	**3.7**	47.2 ± 2.3	48.0 ± 1.2	**-0.8**	43.2 ± 1.3	41.3 ± 0.9	**1.9**
SAT [27]	44.4 ± 1.4	39.3 ± 1.0	5.1	48.0 ± 1.1	47.7 ± 1.4	0.3	44.4 ± 1.5	41.5 ± 0.8	2.9
FC [54]	46.4 ± 1.2	37.8 ± 0.9	8.6	48.7 ± 1.9	45.8 ± 1.3	2.9	46.8 ± 1.4	40.4 ± 0.8	6.4
Att2in [54]	45.9 ± 1.1	38.3 ± 1.0	7.6	47.8 ± 2.0	46.7 ± 1.4	1.1	45.9 ± 1.2	40.9 ± 0.9	5.0
UpDn [4]	48.0 ± 1.3	39.0 ± 0.9	9.0	52.0 ± 1.0	47.3 ± 1.4	4.7	48.5 ± 1.0	41.5 ± 0.9	7.0
Transformer [42]	48.4 ± 0.8	39.7 ± 0.9	8.7	54.1 ± 1.2	48.2 ± 1.1	5.9	47.7 ± 1.2	42.2 ± 0.9	5.5
OSCAR [5]	48.5 ± 1.5	39.3 ± 0.8	9.2	52.5 ± 1.8	47.6 ± 1.2	4.9	48.1 ± 1.1	41.1 ± 0.9	7.0
NIC+ [3]	46.7 ± 1.2	39.5 ± 0.6	7.2	49.5 ± 1.4	47.7 ± 1.5	1.8	46.4 ± 1.2	41.0 ± 0.9	5.4
NIC+Equalizer [3]	51.3 ± 0.7	39.5 ± 0.9	**11.8**	54.8 ± 1.1	47.5 ± 1.4	**7.3**	49.5 ± 0.7	40.9 ± 0.9	**8.6**

racy, which provides some hints about the relationship between accuracy and bias amplification (*Observation 1.4*).

Observation 1.2. **Bias metrics are not consistent.** As analyzed in Section 2.3, different metrics measure different aspects of the bias, so it is expected to produce different results, which may lead to different conclusions. Nevertheless, all the models show bias in all the metrics except Ratio (Table 2.1). However, the relationship between the bias and the models presents different tendencies. For instance, NIC+Equalizer shows the largest bias in LIC (*Observation 1.3*) while Att2in has the largest bias in DBA_O.

Observation 1.3. **NIC+Equalizer increases gender bias with respect to the baseline.** One of the most surprising findings is that even though NIC+Equalizer successfully mitigates gender misclassification when compared to the baseline NIC+ (Error: $12.9 \rightarrow 7.7$ in Table 2.1), it actually increases gender amplification bias by $+4.6$ in LIC. This unwanted side-effect may be produced by the efforts of predicting gender correctly according to the image. As shown in Figure 2.1, NIC+Equalizer tends to produce words that, conversely, are strongly associated with that gender, even if they are not in the image. Results on DBA_O support this reasoning,

revealing that given a gender, NIC+Equalizer rather produces words correlated with that gender.

Observation 1.4. **LIC tends to increase with BLEU-4, and decrease with vocabulary size.** Figure 2.3 shows that larger the vocabulary, the lower the LIC score. This implies that the variety of the words used in the captions is important to suppress gender bias. As per accuracy, we find that the higher the BLEU-4, the larger the bias tends to be. In other words, even though better models produce better captions, they rely on encoded clues that can identify gender.

Observation 1.5. **LIC is robust against encoders.** In Table 2.2, we explore how the selection of language models affects the results of LIC, LIC_M, and LIC_D when using LSTM, BERT-ft, and BERT-pre encoders. Although BERT is known to contain gender bias itself, the tendency is maintained within the three language models: NIC shows the least bias, whereas NIC+Equalizer shows the most.

2.5.3 Quantification of racial bias

Results for racial bias when using LSTM as encoder are reported in Table 2.3, leading to the following observations.

Observation 2.1. **All the models amplify racial bias.** As with gender, all models exhibits $\text{LIC} > 0$. The magnitude difference of racial bias between the models is smaller than in the case of gender (the standard deviation of LIC among the models is 2.4 for gender and 1.3 for race). This indicates that racial bias is amplified without much regard to the structure or performance of the model. In other words, as all the models exhibit similar tendencies of bias amplification, the problem may not only be on the model structure itself but on how image captioning models are trained.

Observation 2.2. **Racial bias is not as apparent as gender bias.** LIC_M scores in Table 2.3 are consistently smaller than in Table 2.2. The mean of the LIC_M score of all the models is 47.0 for gender and 33.7 for race, which is closer to the random chance.

Observation 2.3. **NIC+Equalizer does not increase racial bias with respect to the baseline.** Unlike for gender bias, NIC+Equalizer does not present more racial bias amplification

Table 2.3: Racial bias scores according to LIC, LIC_M, and LIC_D. Captions are not masked and are encoder with LSTM.

Model	LIC_M	LIC_D	LIC
NIC [2]	33.3 ± 1.9	27.6 ± 1.0	5.7
SAT [27]	31.3 ± 2.3	26.8 ± 0.9	**4.5**
FC [54]	33.6 ± 1.0	26.0 ± 0.8	7.6
Att2in [54]	35.2 ± 2.3	26.6 ± 0.9	**8.6**
UpDn [4]	34.4 ± 2.1	26.6 ± 0.9	7.8
Transformer [42]	33.3 ± 2.3	27.2 ± 0.8	6.1
OSCAR [5]	32.9 ± 1.8	27.0 ± 1.0	5.9
NIC+ [3]	34.9 ± 1.5	27.3 ± 1.2	7.6
NIC+Equalizer [3]	34.5 ± 2.8	27.3 ± 0.8	7.2

than NIC+. This indicates that forcing the model to focus on the human area to predict the correct gender does not negatively affect other protected attributes.

2.5.4 Visual and language contribution to the bias

As image captioning is a multimodal task involving visual and language information, bias can be introduced by the image, the language, or both. Next, we investigate which modality contributes the most to gender bias by analyzing the behavior when using partially masked images.

We define three potential sources of bias: 1) the objects being correlated with the gender [23,39,50], 2) the gender of the person in the image [3], and 3) the language model itself [58,62]. To examine them, we mask different parts of the image accordingly: 1) the object that exhibits the highest correlation with gender according to the BA metric, 2) the person, 3) both of the correlated objects and the person. We analyze SAT [27] and OSCAR [5] as representative models of classical and state-of-the-art captioning, respectively. The details of the experiments

can be found in the appendix. LIC_M scores are shown in Table 2.4.

Observation 3.1. **The contribution of objects to gender bias is minimal.** Results *w/o object* show that masking objects do not considerably mitigate gender bias in the generated captions. Compared to the original LIC_M, the score decreases only -1.5 for SAT, and -2.3 for OSCAR, concluding that objects in the image have little impact to the gender bias in the final caption.

Observation 3.2. **The contribution of people to gender bias is higher than objects.** Results *w/o person* show that by masking people in the images, we can reduce bias significantly compared to when hiding objects, indicating that regions associated with humans are the primary source of gender bias among the contents in the image.

Observation 3.3. **Language models are a major source of gender bias.** Results *w/o both* show that even when the gender-correlated objects and people are removed from the images, the generated captions have a large bias (Δ_{Unbias} is $+12.2$ for SAT, $+14.0$ for OSCAR). This indicates that the language model itself is producing a large portion of the bias. To reduce it, it may not be enough to only focus on the visual content, but efforts should also be focused on the language model. Figure 2.4 shows the generated captions and their bias score when images are partly masked.

2.6 Limitations

In Section 2.3, we analyzed multiple fairness metrics and their limitations when applied to image captioning. We proposed LIC with the aim to overcome these limitations and unify the evaluation of societal bias in image captioning. However, LIC also presents several limitations.

Annotations LIC needs images to be annotated with their protected attribute. Annotations are not only costly, but may also be problematic. For example, the classification of race is controversial and strongly associated with the cultural values of each annotator [33], whereas gender is commonly classified as a binary $\{female, male\}$ attribute, lacking inclusiveness with

Table 2.4: Gender bias results with partially masked images. Δ_{Unbias} shows the difference with respect to a non-biased model ($\text{LIC}_M = 25.0$), and Δ_{Original} with respect to the non-masked case.

Model	Image	LIC_M	Δ_{Unbias}	Δ_{Original}
SAT [27]	Original	44.4 ± 1.4	$+19.4$	0.0
	w/o object	42.9 ± 1.6	$+17.9$	-1.5
	w/o person	39.1 ± 1.4	$+14.1$	-5.3
	w/o both	37.2 ± 0.8	$+12.2$	-7.2
OSCAR [5]	Original	48.5 ± 1.5	$+23.2$	0.0
	w/o object	46.2 ± 1.3	$+21.2$	-2.3
	w/o person	39.7 ± 1.3	$+14.7$	-8.8
	w/o both	39.0 ± 1.5	$+14.0$	-9.5

non-binary and other-gender realities.

Training A classifier needs to be trained to make predictions about the protected attributes. The initialization of the model and the amount of training data may impact on the final results. To mitigate this stochastic effect, we recommended to report results conducted on multiple runs.

Pre-existing bias The language encoder may propagate extra bias into the metric if using pretrained biased models, e.g., word embeddings or BERT. To avoid this, we recommend as much random weight initialization as possible.

2.7 Conclusion

We proposed LIC, a metric to quantify societal bias amplification in image captioning. LIC is built on top of the idea that there should not be differences between how demographic sub-

Figure 2.4: Generated captions and bias scores when images are partly masked. The bias score does not decrease when the object (bicycle) and the person (man) are masked.

groups are described in captions. The existence of a classifier that predicts gender and skin tone from generated captions more accurately than from human captions, indicated that image captioning models amplify gender and racial bias. Surprisingly, the gender equalizer designed for bias mitigation presented the highest gender bias amplification, highlighting the need of a bias amplification metric for image captioning.

Appendix

This supplementary material includes:

- Experimental details (Appendix 2.8).
- List of gender-related words (Appendix 2.9).
- More visual examples (Appendix 2.10).
- Additional results (Appendix 2.11).
- Potential negative impact (Appendix 2.12).

2.8 Experimental details

In this section, we provide the details for the experiments.

2.8.1 LIC metric training details

We evaluate three classifiers for LIC (LSTM, BERT-ft, and BERT-pre). Their details and hyperparameters can be found below. All the classifiers are trained with Adam [63].

- **LSTM.** A two-layer bi-directional LSTM [56] with a fully-connected layer on top. Weights are initialized randomly and training is conducted on the training set for 20 epochs, with learning rate 5×10^{-5}.
- **BERT-ft.** BERT-base [60] Transformer with two fully-connected layers with Leaky ReLU activation on top. All the weights are fine-tuned while training. Training is conducted for 5 epochs with learning rate 1×10^{-5}.
- **BERT-pre.** Same architecture as BERT-ft. Only the last fully-connected layers are fine-tuned, whereas BERT weights are frozen. Training is conducted for 20 epochs with learning rate 5×10^{-5}.

2.8.2 Other metrics details

Details for computing BA, DBA$_G$, and DBA$_O$ metrics.

- **BA.** We use nouns, verbs, adjectives, and adverbs of the top $1,000$ common words in the captions as $\mathcal{L}$ and calculate the co-occurrence of the gender words and the common words in the captions. As [23], we filter the words that are not strongly associated with humans by removing words that do not occur with each gender at least 100 times in the ground-truth captions, leaving a total of 290 words.

- **DBA$_G$ and DBA$_O$.** Let p denote the probability calculated by the (co-)occurrence. The definition of DBA$_G$ and DBA$_O$ [50] is:

$$\text{DBA} = \frac{1}{|\mathcal{L}||\mathcal{A}|} \sum_{a \in \mathcal{A}, l \in \mathcal{L}} y_{al}\Delta_{al} + (1 - y_{al})(-\Delta_{al}) \tag{2.12}$$

$$y_{al} = \mathbb{1}\left[p(a, l) > p(a)p(l)\right] \tag{2.13}$$

$$\Delta_{al} = \begin{cases} \hat{p}(a|l) - p(a|l) & \text{for DBA}_G \\ \hat{p}(l|a) - p(l|a) & \text{for DBA}_O \end{cases} \tag{2.14}$$

  For DBA$_G$, we use the MSCOCO objects [28] annotated on the images as $\mathcal{L}$ and gender words in the captions as $\mathcal{A}$. For DBA$_O$, we use the MSCOCO objects [28] in the captions as $\mathcal{L}$ and gender annotations [6] as $\mathcal{A}$.

2.8.3 Image masking

Here, we explain how we masked objects and people in the images to estimate the contribution of each modality to the bias.

- **SAT** [27] uses grid-based deep visual features [64] extracted by ResNet [65]. Thus, we directly mask the objects, people, or both in the images using segmentation mask annotations, and feed the images into the captioning model to generate captions.

Table 2.5: Racial bias scores according to LIC, LIC_M, and LIC_D for several image captioning models. Captions are encoder with LSTM, BERT-ft, or BERT-pre. Unbiased model is $\text{LIC}_M = 25$ and $\text{LIC} = 0$.

	LSTM			BERT-ft			BERT-pre		
Model	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC	LIC_M	LIC_D	LIC
NIC [2]	33.3 ± 1.9	27.6 ± 1.0	5.7	37.0 ± 3.0	36.7 ± 1.1	**0.3**	34.7 ± 2.1	33.6 ± 1.2	1.1
SAT [27]	31.3 ± 2.3	26.8 ± 0.9	**4.5**	38.1 ± 2.7	36.5 ± 1.4	1.6	33.9 ± 1.5	33.3 ± 1.3	**0.6**
FC [54]	33.6 ± 1.0	26.0 ± 0.8	7.6	40.4 ± 2.4	36.4 ± 1.6	4.0	36.9 ± 2.2	32.6 ± 1.2	**4.3**
Att2in [54]	35.2 ± 2.3	26.6 ± 0.9	**8.6**	40.4 ± 2.0	36.1 ± 1.2	**4.3**	36.8 ± 1.9	32.7 ± 1.1	4.1
UpDn [4]	34.4 ± 2.1	26.6 ± 0.9	7.8	40.2 ± 1.7	36.9 ± 1.2	3.3	36.5 ± 2.5	33.2 ± 1.2	3.3
Transformer [42]	33.3 ± 2.3	27.2 ± 0.8	6.1	39.4 ± 1.7	37.4 ± 1.3	2.0	36.2 ± 2.2	34.1 ± 1.2	2.1
OSCAR [5]	32.9 ± 1.8	27.0 ± 1.0	5.9	39.4 ± 2.3	36.9 ± 0.9	2.5	35.5 ± 2.5	32.9 ± 1.1	2.6
NIC+ [3]	34.9 ± 1.5	27.3 ± 1.2	7.6	39.5 ± 2.6	37.1 ± 1.3	2.4	36.8 ± 2.4	33.6 ± 1.3	3.2
NIC+Equalizer [3]	34.5 ± 2.8	27.3 ± 0.8	7.2	38.7 ± 3.1	36.6 ± 1.3	2.1	36.0 ± 2.2	33.4 ± 1.4	2.6

- **OSCAR** [5] leverages region-based deep visual features [4] extracted by a Faster-RCNN [66]. Therefore, instead of masking the objects, people, or both in the images, we remove the region-based features whose bounding box overlaps with the ground truth bounding by more than 50 percent.

2.9 List of gender-related words

We list the gender-related words that are replaced with the special token when inputting to gender classifiers: woman, female, lady, mother, girl, aunt, wife, actress, princess, waitress, sister, queen, pregnant, daughter, she, her, hers, herself, *man*, *male*, *father*, *gentleman*, *boy*, *uncle*, *husband*, *actor*, *prince*, *waiter*, *son*, *brother*, *guy*, *emperor*, *dude*, *cowboy*, *he*, *his*, *him*, *himself* and their plurals. Orange/*Olive* denotes feminine/masculine words used to calculate Ratio, Error, BA, and DBA_G.

2.10 Visual examples

Here, we show more visual examples that could not be included in the main paper due to space limitations. Figure 2.5 shows generated captions and their bias score for all the models evaluated in the main paper. Additionally, Figure 2.6 shows more examples where NIC+Equalizer produces words strongly associated with gender stereotypes even when the evidence is not contained in the image. Whereas in the main paper we showed samples for women, here we show samples for men. It can be seen that NIC+Equalizer generates male-related words (e.g., *suit*, *tie*), and thus, obtain a higher bias score. We also show additional examples when images are partly masked in Figure 2.7. The generated caption when the person (man) and the most correlated object (bicycle) are masked still contains a large bias score towards male.

2.11 Additional results

We compare LIC for race when using different language encoders in Table 2.5. As with gender bias, the results show that LIC is consistent across different language models.

2.12 Potential negative impact

A potential negative impact of the use of the LIC metric to evaluate societal bias in image captioning is that researchers and computer vision practitioners may underestimate the bias and their impact in their models. Although it is important to have a tool to measure societal bias in computer vision models, we need to note that none metric can ensure the actual amount of bias. In other words, even if LIC (or any other metric) is small, or even zero, the model may still be biased. Therefore, relying on a single metric may overlook the problem.

Additionally, whereas we use the value of LIC as the amount of bias amplification on a model, the definition of bias is different among existing work. As there is no standard definition of bias for image captioning, we should notice that our method is, perhaps, not the most

appropriate one for all the contexts, and researchers should carefully consider which metric to use according to each application.

Figure 2.5: For each caption generated by humans or the models evaluated in the paper, we show our proposed bias score for *female* and *male* attributes. The contribution of each word to the bias score is shown in gray-scale (bold for the word with the highest contribution). Gender related words are masked during training and testing.

Figure 2.6: Measuring gender bias in MSCOCO captions [1]. For each caption generated by humans, NIC [2], or NIC+Equalizer [3], we show our proposed bias score for *female* and *male* attributes. The contribution of each word to the bias score is shown in gray-scale (bold for the word with the highest contribution). Gender related words are masked during training and testing.

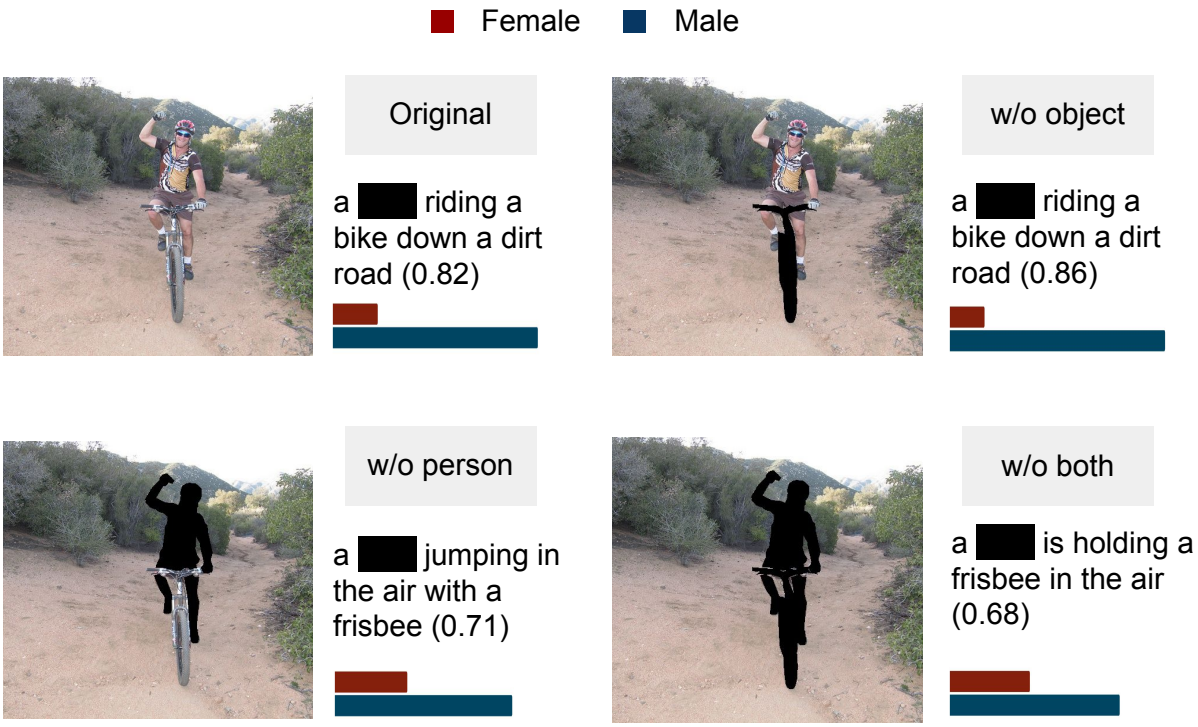


Figure 2.7: Generated captions and bias scores when images are partly masked.

Chapter 3

Model-Agnostic Gender Debiased Image Captioning

3.1 Overview

(a) context → gender bias mitigation (b) gender → context bias mitigation

Figure 3.1: Generated captions by a baseline captioning model (UpDn [4]) and LIBRA. We show the baseline suffers from context → gender/gender → context biases, predicting incorrect gender or incorrect word (e.g., in the left example, *skateboard* highly co-occurs with men in the training set, and the baseline incorrectly predicts *boy*). Our proposed framework successfully modifies those incorrect words.

In computer vision, societal bias, for which a model makes adverse judgments about specific population subgroups usually underrepresented in datasets, is increasingly concerning [22, 24, 33–35, 43, 48, 67–69]. A renowned example is the work by Buolamwini and Gebru [22], which demonstrated that commercial facial recognition models predict Black women with higher error rates than White men. The existence of societal bias in datasets and models is extremely problematic as it inevitably leads to discrimination with potentially harmful consequences against people in already historically discriminated groups.

One of the computer vision tasks in which societal bias is prominent is image captioning [2, 27], which is the task of generating a sentence describing an image. Notably, image captioning models not only reproduce the societal bias in the training datasets, but also amplify it. This phenomenon is known as bias amplification [70–74] and makes models produce sentences more biased than the ones in the original training dataset. As a result, the generated sentences can contain stereotypical words about attributes such as gender that are sometimes irrelevant to the images.

Our study focuses on gender bias in image captioning models. First, based on the observations in previous work [3, 45, 50, 75, 76], we hypothesize that there exist two different types of biases affecting captioning models:

Type 1. context $\rightarrow$ gender bias, which makes captioning models exploit the context of an image and precedently generated words, increasing the probability of predicting certain gender, as shown in Figure 3.1 (a).

Type 2. gender $\rightarrow$ context bias, which increases the probability of generating certain words given the gender of people in an image, as shown in Figure 3.1 (b).

Both types of biases can result in captioning models generating harmful gender-stereotypical sentences.

A seminal method to mitigate gender bias in image captioning is Gender equalizer [3], which forces the model to focus on image regions with a person to predict their gender correctly. Training a captioning model using Gender equalizer successfully reduces gender mis-

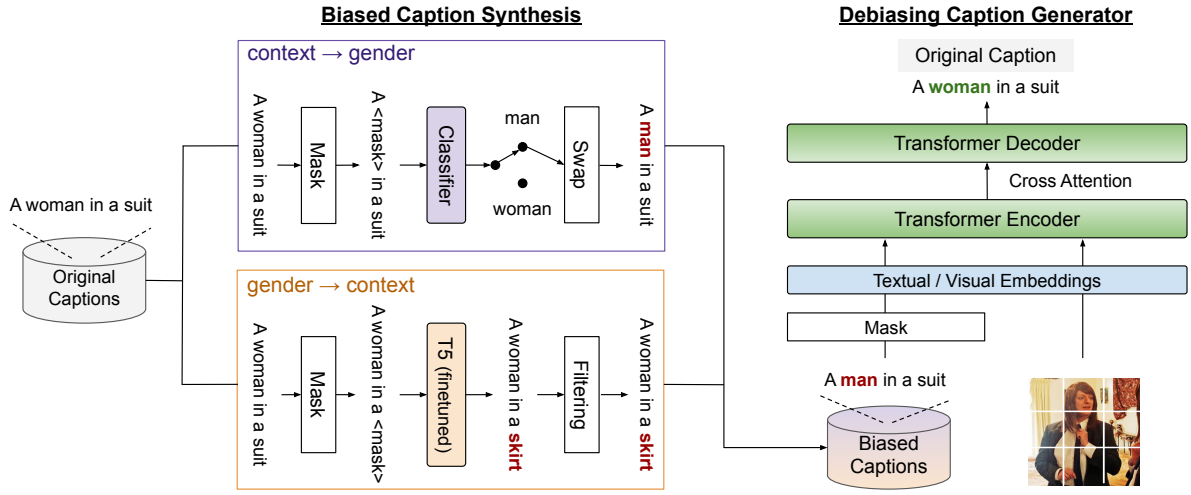


Figure 3.2: Overview of LIBRA. For the original captions (i.e., ground-truth captions written by annotators), we synthesize biased captions with context → gender or/and gender → context bias (Biased Caption Synthesis). Then, given the biased captions and the original images, we train an encoder-decoder captioner, Debiasing Caption Generator, to debias the input biased captions (i.e., predict original captions).

classification (reducing context → gender bias). However, focusing only on decreasing such bias can conversely amplify the other type of bias [50, 75]. For example, as shown in Figure 3.6, a model trained to correctly predict the gender of a person can produce other words that are biased toward that gender (amplifying gender → context bias). This suggests that methods for mitigating bias in captioning models must consider both types of biases.

We propose a method called LIBRA (model-agnostic debiasing framework) to mitigate bias amplification in image captioning by considering both types of biases. Specifically, LIBRA consists of two main modules: 1) Biased Caption Synthesis (BCS), which synthesizes gender-biased captions (Section 3.3), and 2) Debiasing Caption Generator (DCG), which mitigates bias from synthesized captions (Section 3.4). Given captions written by annotators, BCS synthesizes biased captions with gender → context or/and context → gender biases. DCG is then trained to recover the original caption given a ⟨synthetic biased caption, image⟩ pair. Once trained, DCG can be used on top of any image captioning models to mitigate gender bias amplification by

taking the image and generated caption as input. Our framework is model-agnostic and does not require retraining image captioning models.

Extensive experiments and analysis, including quantitative and qualitative results, show that LIBRA reduces both types of gender biases in most image captioning models on various metrics [3, 23, 45, 75]. This means that DCG can correct gender misclassification caused by the context of the image/words that is biased toward a certain gender, mitigating context $\rightarrow$ gender bias (Figure 3.1 (a)). Also, it tends to change words skewed toward each gender to less biased ones, mitigating gender $\rightarrow$ context bias (Figure 3.1 (b)). Furthermore, we show that evaluation of the generated captions' quality by a metric that requires human-written captions as ground-truth (e.g., BLEU [8] and SPICE [9]) likely values captions that imitate how annotators tend to describe the gender (e.g.., *women posing* vs. *men standing*).

3.2 Related work

Societal bias in image captioning In image captioning [4,51], societal bias can come from both the visual and linguistic modalities [3, 6, 50]. In the visual modality, the image datasets used to train captioning models are skewed regarding human attributes such as gender [6,23,32,77,78], in which the number of images with men is twice as much as those of women in MSCOCO [28]. Additionally, captions written by annotators can also be biased toward a certain gender because of gender-stereotypical expressions [6, 76], which can be a source of bias from the linguistic modality. Models trained on such datasets not only reproduce societal bias but amplify it [3, 6, 50, 75]. This phenomenon is demonstrated by Burns et al.. [3], which showed that image captioning models learn the association between gender and objects and make gender distribution in the predictions more skewed than in datasets. We show that LIBRA can mitigate such gender bias amplification in various captioning models. What is better, we demonstrate that our model often produces less gender-stereotypical captions than the original captions.

Mitigating societal bias Mitigation of societal bias has been studied in many tasks [12, 23, 31, 37, 39, 45, 49, 79–82], such as image classification [83] and visual semantic role labeling

[84]. For example, Wang et al.. [39] proposed an adversarial debiasing method to mitigate gender bias amplification in image classification models. In image captioning, Burns et al. [3] proposed the Gender equalizer we described in Section 3.1 to mitigate context → gender bias. However, recent work [50, 75] showed that focusing on mitigating gender misclassification can lead to generating gender-stereotypical words and amplifying gender → context bias. LIBRA is designed to mitigate bias from the two types of biases.

Image caption editing DCG takes a ⟨caption, image⟩ pair as input and debiases the caption. This process is aligned with image caption editing [15, 85, 86] for generating a refined caption. These models aim to correct grammatical errors and unnatural sentences but not to mitigate gender bias. In Section 3.5.3, we compare DCG with a state-of-the-art image caption editing model [15] and show that a dedicated framework for addressing gender bias is necessary.

3.3 Biased caption synthesis

Figure 3.2 shows an overview of LIBRA, consisting of BCS and DCG. This section introduces BCS to synthesize captions with both context → gender or/and gender → context biases.

Notation Let $\mathcal{D} = \{(I, y)\}$ denote a training set of the captioning dataset, where I is an image and $y = (y_1, \ldots, y_N)$ is the ground-truth caption with N tokens. $\mathcal{D}_{\mathrm{g}}$ denotes a subset of $\mathcal{D}$, which is given by filter F_{GW} as

$$\mathcal{D}_{\mathrm{g}} = F_{\mathrm{GW}}(\mathcal{D}), \tag{3.1}$$

F_{GW} keeps captions that contains either women or men words (e.g., *girl*, *boy*).[1] Therefore, samples in $\mathcal{D}_{\mathrm{g}}$ come with a gender attribute $g \in \mathcal{G}$, where $\mathcal{G} = \{\text{female}, \text{male}\}$.[2] We define the set that consists of women and men words as gender words.

[1] We pre-defined women and men words. The list is in the appendix.

[2] In this work, we focus on binary gender categories in our framework and evaluation by following previous work [3, 23]. We recognize that the more inclusive gender categories are preferable, and it is the future work.

3.3.1 Context → gender bias synthesis

Gender-swapping
A **man** with a beer is wearing a gray tie

T5-generation
A woman with a **umbrella** is wearing a gray **shirt**

Merged
A **man** with a **hat** is wearing a gray tie

Original
A woman with a beer is wearing a gray tie

Gender-swapping
A **man** in the water trying to grab a frisbee

T5-generation
A woman in the **kitchen** trying to grab a **pizza**

Merged
A **man** in the **air** trying to grab a frisbee

Original
A woman in the water trying to grab a frisbee

Figure 3.3: Biased captions synthesized by BCS. Gender-swapping denotes synthesized captions by swapping the gender words (Section 3.3.1). T5-generation denotes synthesized captions by T5 (Section 3.3.2). Merged represents biased captions synthesized by applying T5-generation and Gender-swapping (Section 3.3.3).

Context → gender bias means gender prediction is overly contextualized by the image and caption. Therefore, the gender should be predictable from the image and caption context when the caption has context → gender bias. The idea of synthesizing context → gender biased captions is thus to swap the gender words in the original caption to make it consistent with the

context when the gender predicted from the context is skewed toward the other gender. Since an original caption faithfully represents the main content of the corresponding image [87, 88], we can solely use the caption to judge if both image and caption are skewed. To this end, we train a sentence classifier that predicts gender from textual context to synthesize biased captions. We introduce the detailed steps.

Masking Captions with context $\rightarrow$ gender bias are synthesized for $\mathcal{D}_{\mathrm{g}}$. Let F_{PG} denote the filter that removes captions whose gender is predictable by the sentence classifier. Given $y \in \mathcal{D}_{\mathrm{g}}$, F_{PG} instantiated by first masking gender words and replacing corresponding tokens with the mask token to avoid revealing the gender, following [75]. We denote this gender word masking by $\mathrm{mask}(\cdot)$.

Gender classifier We then train[3] gender classifier f_{g} to predict the gender from masked caption as

$$\hat{g} = f_{\mathrm{g}}(y) = \mathrm{argmax}_g \, p(G = g|\mathrm{mask}(y)) \tag{3.2}$$

where $p(G = g|\mathrm{mask}(y))$ is the probability of being gender g given masked y. F_{PG} is then applied to $\mathcal{D}_{\mathrm{g}}$ as:

$$F_{\mathrm{PG}}(\mathcal{D}_{\mathrm{g}}) = \{y \in \mathcal{D}_{\mathrm{g}} | \hat{g}(y) \neq g\}, \tag{3.3}$$

recalling $\hat{g}$ is a function of y.

Gender swapping The inconsistency of context y' and gender g means that y' is skewed toward the other gender; therefore, swapping gender wards (e.g., *man* $\rightarrow$ *woman*) in $y \in F_{\mathrm{PG}}(\mathcal{D}_{\mathrm{g}})$ results in a biased caption. Letting $\mathrm{swap}(\cdot)$ denote this gender swapping operation, the augmenting set $\mathcal{A}_{\mathrm{CG}}$ is given by:

$$\mathcal{A}_{\mathrm{CG}} = \{\mathrm{swap}(y) | y \in F_{\mathrm{PG}}(\mathcal{D}_{\mathrm{g}})\}. \tag{3.4}$$

[3]Refer to the appendix for training details.

Figure 3.3 shows some synthetically biased captions (refer to Gender-swapping). We can see that the incorrect gender correlates with context skewed toward that gender. For instance, in the top example, *tie* is skewed toward men based on the co-occurrence of men words and *tie*.

3.3.2 Gender $\rightarrow$ context bias synthesis

Our idea for synthesizing captions with gender $\rightarrow$ context bias is to sample randomly modified captions of y and keep ones with the bias. Sampling modified captions that potentially suffer from this type of bias is not trivial. We thus borrow the power of a language model. That is, captions with gender $\rightarrow$ context bias tend to contain words that well co-occur with gender words, and this tendency is supposedly encapsulated in a language model trained with a large-scale text corpus. We propose to use the masked token generation capability of T5 [89] to sample modified captions and filter them for selecting biased captions.

T5 masked word generation T5 is one of the state-of-the-art Transformer language models. For better alignment with the vocabulary in the captioning dataset, we finetune T5 with $\mathcal{D}$ by following the process of training the masked language model in [60].[4] After finetuning, we sample randomly modified captions using T5. Specifically, we randomly mask 15% of the tokens in $y \in \mathcal{D}$. Note that we exclude tokens of the gender words if any as they serve as the only cue of the directionality of bias (either men or women).

Let $y_{\mathcal{M}}$ denote a modified y whose m-th token ($m \in \mathcal{M}$) is replaced with the mask token. The masked token generator by T5 can complete the masked tokens solely based on $y_{\mathcal{M}}$, i.e., $\hat{y} = \mathrm{T5}(y_{\mathcal{M}})$. With this, we can sample an arbitrary number of $\hat{y}$'s to make a T5-augmented set $\mathcal{D}_{\mathrm{T5}}$ as[5]:

$$\mathcal{D}_{\mathrm{T5}} = \{\hat{y} = \mathrm{T5}(y_{\mathcal{M}}) | y \in \mathcal{D}, \mathcal{M} \sim \mathcal{R}\}, \tag{3.5}$$

where $\mathcal{M}$ is sampled from set $\mathcal{R}$ of all possible masks.

[4] Refer to the appendix for the details of this finetuning.

[5] We remove trivial modification that replaces a word with its synonyms based on WordNet [90] and unnatural captions with dedicated classifier. More details can be found in the appendix.

Filtering We then apply a filter to $\mathcal{D}_{\text{T5}}$ to remove captions that decrease gender $\rightarrow$ context bias, which is referred to as gender filter. We thus borrow the idea in Eq. (3.3). For this, we only use captions in $\mathcal{D}_{\text{T5}}$ that contain the gender words, i.e., $\mathcal{D}_{\text{T5,g}} = F_{\text{GW}}(\mathcal{D}_{\text{T5}})$, to guarantee that all captions have gender attribute g. To collectively increase gender $\rightarrow$ context bias in the set, we additionally use condition $d(y', y) = p(G = g|\text{mask}(y')) - p(G = g|\text{mask}(y)) > \delta$, which means the gender of $y' \in \mathcal{D}_{\text{T5,g}}$ should be more predictable than the corresponding original $y \in \mathcal{D}_{\text{g}}$ by a predefined margin δ. Gender filter F_{GF} is given by:

$$F_{\text{GF}}(\mathcal{D}_{\text{T5,g}}, \mathcal{D}_{\text{g}}) = \{y' \in \mathcal{D}_{\text{T5,g}} | \hat{g}(y') = g, d(y', y) > \delta\}. \tag{3.6}$$

The appendix shows that F_{GF} can keep more gender-stereotypical sentences than the original captions.

With the gender filter, augmenting set $\mathcal{A}_{\text{GC}}$ is given as the intersection of the filtered sets as:

$$\mathcal{A}_{\text{GC}} = F_{\text{GF}}(\mathcal{D}_{\text{T5,g}}, \mathcal{D}_{\text{g}}). \tag{3.7}$$

As a result, the synthesized captions contain gender-stereotypical words that often co-occur with that gender as shown in Figure 3.3 (refer to T5-generation). For example, in the bottom sample, *kitchen* co-occurs with women words about twice as often as it co-occurs with men words in $\mathcal{D}$, amplifying gender $\rightarrow$ context bias.

3.3.3 Merging together

For further augmenting captions, we merge the processes for augmenting both context $\rightarrow$ gender and gender $\rightarrow$ context biases, which is given by:

$$\mathcal{A} = \{\text{swap}(y) | y \in F_{\text{PG}}(\mathcal{D}_{\text{T5,g}})\}, \tag{3.8}$$

which means that the process for synthesizing context $\rightarrow$ gender bias in Eqs. (3.3) and (3.4) is applied to T5 augmented captions. In this way, the textual context becomes skewed toward the new gender. Some synthesized samples can be found in Figure 3.3 (refer to Merged).

3.4 Debiasing caption generator

DCG is designed to mitigate the two types of gender bias in an input caption to generate a debiased caption.

Architecture DCG has an encoder-decoder architecture. The encoder is a Transformer-based vision-and-language model [91] that takes an image and text as input and outputs a multi-modal representation. The decoder is a Transformer-based language model [92] that generates text given the encoder's output. The encoder's output is fed into the decoder via a cross-attention mechanism [93].

Training Let $\mathcal{D}^{\star} = \mathcal{A}_{\text{CG}} \cup \mathcal{A}_{\text{GC}} \cup \mathcal{A} = \{(I, y^{\star})\}$ denote the set of synthetic biased captions where $y^{\star}$ is a biased caption. When training DCG, given a (I, $y^{\star}$) pair, we first mask 100η percent of words in the input caption. The aim is to add noise to the input sentence so DCG can see the image when refining the input caption, avoiding outputting the input sentence as it is by ignoring the image. The masked caption is embedded to $\bar{y}$ by word embedding and position embedding. The input image I is embedded to $\bar{I}$ through linear projection and position embedding. $\bar{y}$ and $\bar{I}$ are fed into the DCG encoder, and the output representation of the encoder is inputted to the DCG decoder via a cross-attention mechanism. DCG is trained to recover the original caption y with a cross-entropy loss $\mathcal{L}_{\text{ce}}$ as

$$\mathcal{L}_{\text{ce}} = -\sum_{t=1}^{N} \log p(y_t | y_{1:t-1}, I, y^{\star}) \tag{3.9}$$

where p is conditioned on the precedently generated tokens, and I and $y^{\star}$ through the cross-attention from the encoder. The trained DCG learns to mitigate two types of biases that lie in the input-biased captions.

Inference We apply the trained DCG to the output captions of captioning models. Let y_{c} denote a generated caption by an image captioning model. As in training, given a pair of (I, y_{c}), we first mask 100η percent of words in the input caption. Then, DCG takes the masked caption and image and generates a debiased caption. DCG can be used on top of any image captioning

Table 3.1: Dataset construction. Swap denotes synthesized captions by Gender-swapping (Section 3.3.1). T5 denotes synthesized captions by T5-generation (Section 3.3.2). Ratio represents the ratio of the number of each type of biased data.

Synthesis method				
Swap	T5	Merged	Ratio	Num. sample
✓	✓	-	1:1:0	57,284
-	✓	✓	0:1:1	114,568
✓	✓	✓	1:2:1	114,568

models and does not require training in captioning models to mitigate gender bias.

3.5 Experiments

Dataset We use MSCOCO captions [1]. For training captioning models, we use the MSCOCO training set that contains $82,783$ images. For evaluation, we use a subset of the MSCOCO validation set, consisting of $10,780$ images, that come with binary gender annotations from [6]. Each image has five captions from annotators.

For synthesizing biased captions with BCS, we use the MSCOCO training set. The maximum number of synthetic captions by Gender-swapping is capped by $|F_{\mathrm{PG}}(\mathcal{D}_g)| = 28,642$, while T5-generation and Merged can synthesize an arbitrary number of captions by sampling $\mathcal{M}$. We synthesize captions so that the number of captions with gender swapping (i.e., Gender-Swapping and Merged) and T5-generation can be balanced as in Table 3.1.

Bias metrics We mainly rely on three metrics to evaluate our framework: 1) **LIC** [75], which compares two gender classifiers' accuracies trained on either generated captions by a captioning

model or human-written captions. Higher accuracy of the classifier trained on a model's predictions means that the model's captions contain more information to identify the gender in images, indicating gender → context bias amplification, 2) **Error** [3], which measures the gender misclassification ratio of generated captions. We consider Error to evaluate context → gender bias whereas it does not directly measure this bias (discussed in the appendix) , and 3) **BiasAmp** [23], a bias amplification measurement based on word-gender co-occurrence, which is possibly the cause of gender → context bias. More details about these bias metrics are described in the appendix.

Captioning metrics The accuracy of generated captions is evaluated on reference-based metrics that require human-written captions to compute scores, specifically BLEU-4 [8], CIDEr [52], METEOR [10], and SPICE [9]. While those metrics are widely used to evaluate captioning models, they often suffer from disagreements with human judges [7]. Thus, we also use a reference-free metric, CLIPScore [7], that relies on the image-text matching ability of the pre-trained CLIP. CLIPScore has been shown to have a higher agreement with human judgment than reference-based metrics.

Captioning models We evaluate two standard types of captioning models as baselines: 1) CNN encoder-LSTM decoder models (NIC [2], SAT [27], FC [54], Att2in [54], and UpDn [4]) and 2) state-of-the-art Transformer-based models (Transformer [42], OSCAR [5], ClipCap [94], and GRIT [11]). Note that most of the publicly available pre-trained models are trained on the training set of the Karpathy split [57] that uses the training and validation sets of MSCOCO for training. As we use the MSCOCO validation set for our evaluation, we retrain the captioning models on the MSCOCO training set only.

Debiasing methods As debiasing methods, we compare LIBRA against Gender equalizer [3]. Gender equalizer utilizes extra segmentation annotations in MSCOCO [28], which are not always available. The method is not applicable to captioning models that use object-based visual features such as Faster R-CNN [66] because the pre-trained detector's performance drops

GT gender: Male

OSCAR
a **man** flying through the air while riding a skateboard

+LIBRA
a **girl** glides through the air while riding a skateboard

GT gender: Female

OSCAR
a **man** riding a wave on top of a surfboard

+LIBRA
a **woman** riding a wave on top of a surfboard

Figure 3.4: Gender misclassification of LIBRA (Top). Gender misclassification of OSCAR [5] (Bottom). GT gender denotes ground-truth gender annotation in [6].

considerably for human-masked images.[6] In the experiment, we apply Gender equalizer and LIBRA to debias NIC+, which is a variant of NIC with extra training on images of female/male presented in [3].

For LIBRA, we use $\delta = 0.2$. The vision-and-language encoder of DCG is Vilt [91], and the decoder is GPT-2 [92]. Unless otherwise stated, we use the combination of biased data composed of T5-generation and Merged in Table 3.1. We set $\eta = 0.2$ and conduct ablation studies of the settings in Section 3.5.4 and the appendix.

3.5.1 Bias mitigation analysis

We apply LIBRA on top of all the captioning models to evaluate if it mitigates the two types of gender biases. We also report caption evaluation scores based on captioning metrics. Results are shown in Table 3.2. We summarize the main observations as follows:

[6]Faster-RCNN mAP drops from 0.41 to 0.37, and for the person class recall drops from 0.79 to 0.68.

References
- A man is standing next to his ice cream truck
- A man standing in front of a white ice cream truck
- A man is standing in front of his ice cream vehicle
- A ice cream truck parked on the side of the road with the driver standing beside it
- A man is standing next to an ice cream truck

Baseline
A man **standing** in front of a white truck

BLEU-4: 79.6 ↑ METEOR: 43.8 ↑
SPICE: 42.9 ↑ CLIPScore: 74.6 ↓

+LIBRA
A man **posing** in front of a white truck

BLEU-4: 47.5 ↓ METEOR: 33.7 ↓
SPICE: 30.8 ↓ CLIPScore: 76.7 ↑

Figure 3.5: CLIPScore [7] vs. reference-based metrics [8–10]. References denote the ground-truth captions written by annotators. Bold words in the generated captions mean the difference between baseline and LIBRA. Highlighted words in references denote the words that match the bold word in the baseline. We can see that CLIPScore is more robust against word-changing.

LIBRA mitigates gender → context bias. The results on LIC show that applying LIBRA consistently decreases gender → context bias in all the models. We show some examples of LIBRA mitigating bias in Figure 3.1 (b). For example, in the second sample from the right, the baseline, UpDn [4], produces the incorrect word, *suit*. The word *suit* is skewed toward men, co-occurring with men 82% of the time in the MSCOCO training set. LIBRA changes *suit* to *jacket*, mitigating gender → context bias. Besides, in some cases where LIC is negative (i.e., NIC, SAT, FC, Att2in, and ClipCap), the gender → context bias in the generated captions by LIBRA is less than those of human annotators. In the appendix, we show some examples that LIBRA generates less biased captions than annotators' captions.

The results of BiasAmp, which LIBRA consistently reduces, show that LIBRA tends to equalize the skewed word-gender co-occurrences. For example, LIBRA mitigates the co-occurrence of the word *little* and women from 91% in captions by OSCAR to 60%. Results on BiasAmp support the effectiveness of LIBRA regarding the ability to mitigate gender → context

bias.

LIBRA mitigates context → gender bias in most models. The Error scores show that LIBRA reduces gender misclassification in most models except for OSCAR and GRIT ($3.0 \rightarrow 4.6$ for OSCAR, $3.5 \rightarrow 4.1$ for GRIT). We investigate the error cases when LIBRA is applied to OSCAR and find that gender misclassification of LIBRA is often caused by insufficient evidence to identify a person's gender. For instance, in the top example in Figure 3.4, the ground-truth gender annotation is *male*, and OSCAR generates *man* although the person is not pictured properly enough to determine gender.[7] This may suggest that OSCAR learns to guess the gender based on the context, in this case, skateboard[8] to increase gender classification accuracy. However, this causes context → gender bias for images with a gender-context combination rarely seen in the dataset (e.g., women-surfing). In Figure 3.4 (bottom), OSCAR predicts incorrect gender for the image with a male-biased context.[9] In the appendix, we discuss possible solutions for reducing gender misclassification without relying on the context.

LIBRA is good at CLIPScore. The results of the captioning metrics show that CLIPScore is better or almost as high as the baselines when applying LIBRA. As CLIPScore is based on an image-caption matching score, we can confirm that LIBRA does not generate less biased sentences by producing irrelevant words to images. This observation verifies that applying LIBRA on top of the captioning models does not hurt the quality of captions.

CLIPScore versus other metrics. While LIBRA works well on CLIPScore, the score in the reference-based metrics decreases for some models. We examine the cause of the inconsistency between CLIPScore and reference-based metrics and find that generating words that reduce bias hurts reference-based metrics. We show an example in Figure 3.5. LIBRA changes *standing* to *posing*, which is also a valid description of the image. However, the scores of reference-based metrics substantially drop (e.g., $79.6 \rightarrow 47.5$ in BLEU-4). Human annotators tend to

[7] Previous work [76] has shown human annotators possibly annotate gender from context for images without enough cues to judge gender.

[8] *Skateboard* is highly skewed toward men in the dataset, which co-occurs with men more than 90%.

[9] *Surfboard* highly co-occur with men in MSCOCO.

Original
a **man** and a baby elephant standing in the water

+Equalizer
a **woman** in a **bikini** standing next to a dog

+LIBRA
a **woman** and a baby elephant standing in the sand

Original
a man and a **woman** standing next to each other

+Equalizer
a man in a **suit** is holding a laptop

+LIBRA
a man and a **child** standing next to each other

Figure 3.6: LIBRA vs. Gender equalizer [3].

use *posing* for women.[10] Therefore, reference-based metrics value captions that imitate how annotators describe each gender. On the other hand, LIBRA tends to change words skewed toward each gender to more neutral ones, which can be the cause of decreasing scores in the reference-based metrics.

3.5.2 Comparison with other bias mitigation

We compare the performance of LIBRA and Gender equalizer [3] on NIC+ [2], following the code provided by the authors. The results are shown in Table 3.3. As reported in previous work [50, 75], Gender equalizer amplifies gender $\rightarrow$ context bias ($1.4 \rightarrow 6.8$ in LIC) while mitigating gender misclassification ($14.6 \rightarrow 7.8$ in Error). In contrast, LIBRA mitigates gender $\rightarrow$ context and context $\rightarrow$ gender biases, specifically $1.4 \rightarrow 0.4$ in LIC and $14.6 \rightarrow 5.1$ in Error. In the upper sample of Figure 3.6, LIBRA predicts the correct gender while not generating gender-stereotypical words. The results of the comparison with Gender equalizer highlight

[10]The co-occurrence of women and *posing* is more than 60% of the time in the MSCOCO training set.

the importance of considering two types of biases for gender bias mitigation.

3.5.3 Comparison with image caption editing model

We compare LIBRA with a state-of-the-art image caption editing model [15] (refer to ENT). Specifically, we apply LIBRA and ENT on top of the various captioning models and evaluate them in terms of bias metrics and captioning metrics. We re-train ENT by using the captions from SAT [27] for textual features. The results for OSCAR [5] are shown in Table 3.4. The complete results are in the appendix. As for LIC, while LIBRA consistently mitigates gender $\rightarrow$ context bias, ENT can amplify the bias in some baselines (SAT, Att2in, OSCAR, ClipCap, GRIT). Regarding Error, LIBRA outperforms in most baselines except for OSCAR and GRIT. From these observations, we conclude that a dedicated framework for addressing gender bias is necessary to mitigate gender bias.

3.5.4 Ablations

We conduct ablation studies to analyze the influence of different settings of LIBRA. Here, we show the results when applying LIBRA to UpDn [4] and OSCAR [5]. The complete results of all the baselines are in the appendix.

Combinations of synthetic data We compare the performance of the different dataset combinations for training DCG in Table 3.1. The results are shown in Table 3.5. The Error score of T5-generation and Merged is consistently the best among the combinations. As for LIC, the results are not as consistent, but still DCG trained on all types of combinations decreases the score. We chose T5-generation and Merged as it well balances LIC and Error.

Synthetic data evaluation To demonstrate the effectiveness of BCS, we compare LIBRA and DCG trained on captions with random perturbation, which does not necessarily increase gender bias. In order to synthesize such captions, we randomly mask 15 percent of the tokens in the original captions in $\mathcal{D}_{\mathrm{g}}$ and generate words by T5, but without using any filters in Section

3.3. When selecting masked tokens, we allow choosing gender words so that T5 can randomly change the gender. As a result, the synthesized captions contain incorrect words, which are not necessarily due to gender bias. We show the results in Table 3.6. Using biased samples from BCS to train DCG consistently produces the best results in LIC and Error. From this, we conclude that BCS, which intentionally synthesizes captions with gender biases, contributes to mitigating gender biases.

3.6 Conclusion

LIBRA[11] is a model-agnostic framework to mitigate both context $\rightarrow$ gender and gender $\rightarrow$ context biases in captioning models. We experimentally showed that LIBRA mitigates gender bias in multiple captioning models, correcting gender misclassification caused by context and changing to less gender-stereotypical words. To do this, LIBRA synthesizes biased captions using a language model and filtering for intentionally increasing gender biases. Interestingly, the results showed these synthetic captions are a good proxy of gender-biased captions from various captioning models and facilitate model-agnostic bias mitigation. As future work, we will use LIBRA to mitigate other types of bias, such as age or skin-tone, which requires specific annotations, such as the ones in concurrent work [78], and mechanisms to identify each type of bias.

[11]This work is partly supported by JST CREST Grant No. JPMJCR20D3, JST FOREST Grant No. JPMJFR216O, JSPS KAKENHI No. JP22K12091, and Grant-in-Aid for Scientific Research (A).

Table 3.2: Gender bias and captioning quality for several image captioning models. Green/red denotes LIBRA mitigates/amplifies bias with respect to the baselines. For bias, lower is better. For captioning quality, higher is better. LIC and BiasAmp are scaled by 100. Note that CLIPScore for ClipCap can be higher because CLIPScore and ClipCap use CLIP [14] in their frameworks.

	Gender bias ↓			Captioning quality ↑				
Model	LIC	Error	BiasAmp	BLEU-4	CIDEr	METEOR	SPICE	CLIPScore
NIC [2]	0.5	23.6	1.61	21.9	58.3	21.6	13.4	65.2
+LIBRA	-0.3	5.7	-1.47	24.6	72.0	24.2	16.5	71.7
SAT [27]	-0.3	9.1	0.92	34.5	94.6	27.3	19.2	72.1
+LIBRA	-1.4	3.9	-0.48	34.6	95.9	27.8	20.0	73.6
FC [54]	2.9	10.3	3.97	32.2	94.2	26.1	18.3	70.0
+LIBRA	-0.2	4.3	-1.11	32.8	95.9	27.3	19.7	72.9
Att2in [54]	1.1	5.4	-1.01	36.7	102.8	28.4	20.2	72.6
+LIBRA	-0.3	4.6	-3.39	35.9	101.7	28.5	20.6	73.8
UpDn [4]	4.7	5.6	1.46	39.4	115.1	29.8	22.0	73.8
+LIBRA	1.5	4.5	-2.23	37.7	110.1	29.6	22.0	74.6
Transformer [42]	5.4	6.9	0.09	35.0	101.5	28.9	21.1	75.3
+LIBRA	2.3	5.0	-0.26	33.9	98.7	28.6	20.9	75.7
OSCAR [5]	2.4	3.0	1.78	39.4	119.8	32.1	24.0	75.8
+LIBRA	0.3	4.6	-1.95	37.2	113.1	31.1	23.2	75.7
ClipCap [94]	1.1	5.6	1.51	34.8	103.7	29.6	21.5	76.6
+LIBRA	-1.5	4.5	-0.57	33.8	100.6	29.3	21.4	76.0
GRIT [11]	3.1	3.5	3.05	42.9	123.3	31.5	23.4	76.2
+LIBRA	0.7	4.1	1.57	40.5	116.8	30.6	22.6	75.9

Table 3.3: Comparison with Gender equalizer [3]. Green/red denotes the bias mitigation method mitigates/amplifies bias.

	Gender bias ↓		Captioning quality ↑	
Model	LIC	Error	SPICE	CLIPScore
NIC+ [2]	1.4	14.6	17.5	69.9
+Equalizer [3]	6.8	7.8	16.8	69.9
+LIBRA	0.4	5.1	18.9	72.7

Table 3.4: Comparison with image caption editing model. Bold numbers represent the best scores in ENT [15] or LIBRA.

	Gender bias ↓		Captioning quality ↑	
Model	LIC	Error	SPICE	CLIPScore
OSCAR [5]	2.4	3.0	24.0	75.8
+ENT [15]	5.7	**2.8**	21.9	72.8
+LIBRA	**0.3**	4.6	23.2	75.7

Table 3.5: Comparison of data used for training DCG. Bold numbers denote the best scores among the types of synthetic datasets.

	Synthesis method			Gender bias ↓	
Model	Swap	T5	Merged	LIC	Error
UpDn [4]	-	-	-	4.7	5.6
+LIBRA	✓	✓	-	2.3	6.2
+LIBRA	-	✓	✓	1.5	**4.5**
+LIBRA	✓	✓	✓	**1.1**	5.2
OSCAR [5]	-	-	-	2.4	3.0
+LIBRA	✓	✓	-	**-0.8**	6.8
+LIBRA	-	✓	✓	0.3	**4.6**
+LIBRA	✓	✓	✓	0	5.0

Table 3.6: Comparison with random perturbation. Rand. pert. denotes DCG trained on data with random perturbation. Bold numbers denote the best scores in the DCG trained on either biased captions from BCS or captions with random perturbation.

	Gender bias ↓		Captioning quality ↑	
Model	LIC	Error	SPICE	CLIPScore
UpDn [4]	4.7	5.6	22.0	73.8
+Rand. pert.	2.2	5.9	21.8	74.4
+LIBRA	**1.5**	**4.5**	22.0	74.6
OSCAR [5]	2.4	3.0	24.0	75.8
+Rand. pert.	2.0	5.6	22.9	75.4
+LIBRA	**0.3**	**4.6**	23.2	75.7

Appendix

3.7 Details of BCS

In this section, we provide the details for BCS.

3.7.1 Training gender classifier

The gender classifier f_{g} is trained on $\mathcal{D}_{\mathrm{g}}$. Specifically, following [75], we split the captions in $\mathcal{D}_{\mathrm{g}}$ into a balanced split with an equal number of samples per female/male, having $66,526$ captions. We use BERT-base [60] with two fully-connected layers with Leaky ReLU as f_{g} and finetune it on the balanced split. The learning rate is 1×10^{-5}, and the training is conducted on 5 epochs.

3.7.2 Finetuning T5

Following the masked language model in [60], we finetune T5 on captions in $\mathcal{D}$. Specifically, we mask 10% of the tokens in the original caption y. Given the masked caption and the positions of masked tokens $\mathcal{M} = \{m_1, \ldots, m_{|\mathcal{M}|}\}$, T5 predicts the probability of masked tokens by $\prod_{m \in \mathcal{M}} p(y_m | y_{\setminus \mathcal{M}})$, where $y_{\setminus \mathcal{M}}$ denotes all words in an input caption y except for masked tokens $\{y_m\}$. The sample-wise loss is defined as:

$$\mathcal{L}_{mlm} = -\log \prod_{m \in \mathcal{M}} p(y_m | y_{\setminus \mathcal{M}}) \tag{3.10}$$

where $p(y_m | y_{\setminus \mathcal{M}})$ is the output probability of masked token y_m given $y_{\setminus \mathcal{M}}$ from T5.

3.7.3 Details of T5 masked word generation

To remove trivial modifications, we avoid generating synonyms of the masked tokens by using WordNet [90]. When selecting masked tokens, we chose nouns/verbs/adjectives/adverbs based on POS tagging with NLTK [95].

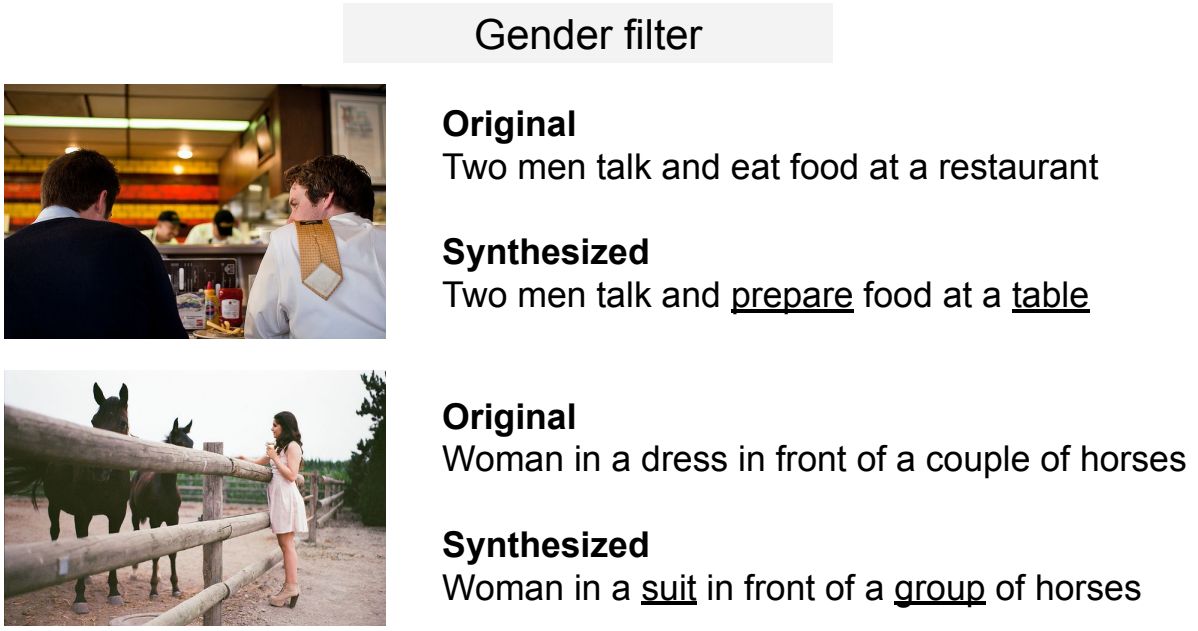


Figure 3.7: Synthesized captions that are removed by the gender filter.

We apply a filter to remove unnatural captions, called an authenticity filter. The authenticity filter uses a classifier that predicts whether an input sentence is synthetic or authentic. Specifically, we train classifier f_{a} with $\mathcal{D}_{\mathrm{T5,g}} \cup \mathcal{D}_g$ to predict whether y is from $\mathcal{D}_{\mathrm{T5,g}}$ or $\mathcal{D}_g$. Let $b \in \{\mathrm{syn}, \mathrm{auth}\}$, prediction $\hat{b}$ is given by:

$$\hat{b} = f_{\mathrm{a}}(y) = \mathrm{argmax}_b \, p(B = b|y) \tag{3.11}$$

where $p(B = b|y)$ is a confidence score that an input caption is b. Thus, if $p(B = \mathrm{authy}|y)$ is close to 1, y is likely to be authentic. We use this classifier to filter less natural captions, i.e.,

$$F_{\mathrm{AF}}(\mathcal{D}_{\mathrm{T5,g}}) = \{y \in \mathcal{D}_{\mathrm{T5,g}} | p(B = \mathrm{auth}|y) > \alpha)\}, \tag{3.12}$$

where α is a predefined threshold. We set $\alpha = 0.3$ and use the same classifier as f_{g} for f_{a}.

3.7.4 Examples of gender/authenticity filter

In Figure 3.7, we show some synthesized captions that are filtered out by the gender filter. The removed captions do not increase gender bias with respect to the original captions. For instance,

Table 3.7: Synthesized captions that are passed or removed by the authenticity filter

Original	Passed	Removed
Woman is sitting near a red train	Woman is sitting near a passenger train	Woman sits sitting near a red train
A man wearing glasses, suit, and tie	A man wearing sunglasses, hat, and tie	A man wearing glasses, glasses, and tie
A man fixing the inside of a toilet	A man fixing the inside of a kitchen	A man holding the inside of a toilet
Women are playing a video game	Women are playing a baseball game	Women are playing a video show

in the bottom example, the word *dress* which is skewed toward women is replaced with *suit* which is skewed toward men. Thus the synthesized caption reduces gender bias compared to the original caption, and it is filtered out by the gender filter.

In Table 3.7, we show some synthesized captions that are passed or removed by the authenticity filter. The examples show that the authenticity filter removes unnatural-sounding or grammatically incorrect captions.

3.8 Details of bias metrics

BiasAmp As for BiasAmp, we also follow the settings presented in [75]. To compute gender-word co-occurrences, we use the top $1,000$ frequent words in $\mathcal{D}$. Specifically, we select nouns / verbs / adjectives / adverbs in the top $1,000$ words. Following [23], we use words that are strongly related to humans by removing words that do not appear more than 100 times with women/men words.

3.9 Additional experiments

3.9.1 Comparison with image caption editing model

We compare LIBRA with a state-of-the-art image caption editing model [15] (refer to ENT). Specifically, we apply LIBRA and ENT on top of the various captioning models and evaluate them in terms of bias metrics and captioning metrics. We re-train ENT by using the captions

from SAT [27] for textual features.[12] The results are shown in Table 3.8. As for LIC, while LIBRA consistently mitigates gender → context bias, ENT can amplify the bias in some baselines (SAT, Att2in, OSCAR, ClipCap, GRIT). Regarding Error, LIBRA outperforms in most baselines except for OSCAR and GRIT. From these observations, we conclude that a dedicated framework for addressing gender bias is necessary to mitigate gender bias.

3.9.2 Analysis of masking

We evaluate the effectiveness of masking input captions in DCG. Specifically, we compare LIBRA with DCG whose input captions are not masked (i.e., $\eta = 0$). The results are shown in Table 3.9. We can see that masking the input captions of DCG consistently improves the scores on bias metrics, which contributes to mitigating two types of biases.

3.9.3 Complete results of ablations

Here, we show the complete results of the ablations in the main paper.

Combinations of synthetic data The complete results of all the baselines are shown in Table 3.10. As in the analysis of the main paper, the results of LIC are not as consistent while DCG trained on all types of combinations mitigate gender → context bias. Regarding Error, DCG trained on T5-generation and Merged has the best results.

Synthetic data evaluation Table 3.11 shows the results of the comparison with random perturbation. This extended table also shows that biased samples from BCS to train DCG produces the best results in LIC and Error in most baselines, which shows the effectiveness of BCS in mitigating gender bias.

[12] In the original paper, the authors use the captions from AoANet [96]. We use SAT for training ENT as AoANet is trained on Karpathy split [57].

3.10 More visual examples

Bias mitigation by LIBRA Figure 3.8 shows the additional examples that LIBRA mitigates context → gender or gender → context bias. For instance, in the left example of (a), the word *motorcycle* highly co-occurs with men in the MSCOCO training set,[13] and the baseline predicts the incorrect gender *man* probably due to context → gender bias. Applying LIBRA on top of the baseline results in mitigating that bias by predicting the correct gender.

Synthesized captions by BCS In Figure 3.9, we show some additional examples of the synthesized captions by BCS. The synthesized captions contain context → gender or/and gender → context biases.

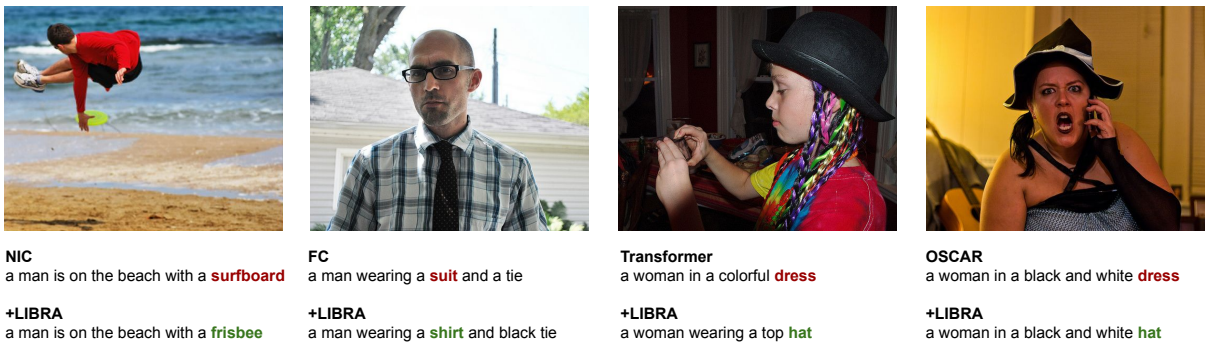
LIBRA vs. human captions The experimental results in the main paper show that LIBRA generates less biased captions than human annotations, resulting in negative LIC scores. Figure 3.10 shows some visual examples that LIBRA generates more neutral words than human captions. For instance, in the left sample, both human and baseline captions contain *short skirt*, which is women's stereotypical words while LIBRA uses more neutral words *tennis outfit*.

Error cases of LIBRA vs. state-of-the-art models In Figure 3.11, we show the additional examples of the error cases of LIBRA and the state-of-the-art models, OSCAR [5] and GRIT [11]. As in the explanation in the main paper, state-of-the-art models can guess gender from the context when there is no clear evidence to identify gender, which leads to amplify context → gender bias.

CLIPScore vs. reference-based metrics In Figure 3.12, we show the additional examples that LIBRA hurts reference-based metrics by generating words that reduce bias whereas LIBRA does not hurt CLIPScore [7]. For instance, in the left example, the word *little* is skewed toward

[13] Co-occurrence of *Motorcycle* and men is about 2.7 times the co-occurrence of *Motorcycle* and women.

(a) context → gender bias mitigation



(b) gender → context bias mitigation

Figure 3.8: Generated captions by the baseline captioning models and LIBRA. We show the baseline suffers from context → gender/gender → context biases, predicting incorrect gender or incorrect word. Our proposed framework successfully modifies those incorrect words.

women in the training set, and LIBRA changed it to *young* which is the less biased word.[14] Both captions correctly describe the image, but LIBRA degrades the scores for the reference-based metrics as human annotators tend to use *little* for women. On the other hand, CLIPScore is more robust against such word-changing.

3.11 List of gender words

The gender words that consist of women and men words are as below:

[14]The co-occurrence of women and *little* is more than 70% of the time in the MSCOCO training set, while *young* is balanced between the gender.

Figure 3.9: Biased captions synthesized by BCS.

Figure 3.10: Comparison of captions from human annotators, baseline, and LIBRA.

woman, female, lady, mother, girl, aunt, wife, actress, princess, waitress, daughter, sister, queen, chairwoman, policewoman, girlfriend, pregnant, daughter, she, her, hers, herself, *man*, *male*, *father*, *gentleman*, *boy*, *uncle*, *husband*, *actor*, *prince*, *waiter*, *son*, *brother*, *guy*, *emperor*, *dude*, *cowboy*, boyfriend, chairman, policeman, *he*, *his*, *him*, *himself* and their plurals. Orange/*Olive* denote women / men words, respectively.

3.12 Limitations

While LIBRA shows superior performance in mitigating gender bias, it also presents some limitations.

Attributes other than gender Gender tends to be described in captions. However, other types of societal biases such as racial bias may not appear as explicitly mentioned in the text and tend to be more subtle, for which our bias mitigation method may not work properly.

GT gender: Female

OSCAR
a couple of **women** standing in a kitchen

+LIBRA
a couple of **men** standing in a kitchen

GT gender: Male

GRIT
a **man** wearing a black jacket and a tie

+LIBRA
a **woman** wearing a black jacket and a tie

GT gender: Male

OSCAR
a **woman** sitting on a bed with a baby and a dog

+LIBRA
a **man** sitting on a bed with a baby and a dog

GT gender: Female

GRIT
a young **boy** riding a wave on a surfboard

+LIBRA
a young **girl** riding a wave on a surfboard

Figure 3.11: Gender misclassification of LIBRA (Top). Gender misclassification of OSCAR [5] or GRIT [11] (Bottom). GT gender denotes ground-truth gender annotation in [6].

References
- The little girl is standing between the low shrub and the fire plug
- A young person standing next to a red fire hydrant
- A little girl leaning against a red fire hydrant.
- A cute little girl standing next to a red fire hydrant.
- A little girl with a stamp on her hand stands beside a fire hydrant.

Baseline
A **little** girl standing next to a red fire hydrant
BLEU-4: 96.2 ↑ METEOR: 51.1 ↑
SPICE: 45.2 ≡ CLIPScore: 98.8 ↓

+LIBRA
A **young** girl standing next to a red fire hydrant
BLEU-4: 83.1 ↓ METEOR: 48.6 ↓
SPICE: 45.2 ≡ CLIPScore: 99.7 ↑

References
- Two men who are standing near each other.
- The two men wearing suits are posing for a picture.
- Two men in suits are pictured standing and smiling.
- A couple of men in ties and suits with glasses on.
- Two men wearing glasses and wearing suits.

Baseline
Two men in **suits** standing next to each other
BLEU-4: 36.9 ↑ METEOR: 41.2 ≡
SPICE: 38.1 ↑ CLIPScore: 80.0 ↓

+LIBRA
Two men in **glasses** standing next to each other
BLEU-4: 0.0 ↓ METEOR: 41.2 ≡
SPICE: 28.6 ↓ CLIPScore: 80.8 ↑

Figure 3.12: CLIPScore [7] vs. reference-based metrics [8–10]. References denote the ground-truth captions written by annotators. Bold words in the generated captions mean the difference between baseline and LIBRA. Highlighted words in references denote the words that match the bold word in the baseline. We can see that CLIPScore is more robust against word changing.

Error for measuring context → gender bias Even though Error can measure one of the aspects of context → gender bias where models make an incorrect prediction of gender based on the context, it does not directly evaluate this bias as it can also occur when predictions are correct but based on the context. Thus, a metric dedicated to context → gender bias would be more insightful.

Predicting gender-neutral words In Section 5.1 in the main paper, we showed that gender misclassification by LIBRA is likely to be caused by the deficient clues to judge gender. A possible solution to mitigate such misclassification without exploiting contextual cues would

be to force the model to predict gender-neutral words such as *person* when there is not enough information to judge gender. We leave this extension as future work.

3.13 Potential negative impact

While LIBRA mitigates gender bias in the bias metrics, it does not ensure that LIBRA completely removes bias. In other words, even though LIBRA works on the bias metrics, the captioning models can still be biased. Thus, a potential negative impact of the use of LIBRA to mitigate gender bias is that the users of LIBRA may become overly confident that LIBRA eliminates gender bias and overlook the problem of gender bias in captioning models. We should carefully consider gender bias in image captioning as it can also exist in aspects not measured by existing metrics.

Table 3.8: Comparison with image caption editing models. Bold numbers represent the best scores in ENT [15] or LIBRA.

	Gender bias ↓		Captioning quality ↑				
Model	LIC	Error	BLEU-4	CIDEr	METEOR	SPICE	CLIPScore
NIC [2]	0.5	23.6	21.9	58.3	21.6	13.4	65.2
+ENT [15]	**-0.3**	22.5	25.8	67.7	22.5	14.3	65.3
+LIBRA	**-0.3**	**5.7**	24.6	72.0	24.2	16.5	71.7
SAT [27]	-0.3	9.1	34.5	94.6	27.3	19.2	72.1
+ENT [15]	1.6	9.9	35.3	96.3	27.3	19.2	71.1
+LIBRA	**-1.4**	**3.9**	34.6	95.9	27.8	20.0	73.6
FC [54]	2.9	10.3	32.2	94.2	26.1	18.3	70.0
+ENT [15]	1.7	10.3	32.9	92.0	26.2	18.2	69.2
+LIBRA	**-0.2**	**4.3**	32.8	95.9	27.3	19.7	72.9
Att2in [54]	1.1	5.4	36.7	102.8	28.4	20.2	72.6
+ENT [15]	2.8	5.3	37.4	103.2	28.4	20.3	71.6
+LIBRA	**-0.3**	**4.6**	35.9	101.7	28.5	20.6	73.8
UpDn [4]	4.7	5.6	39.4	115.1	29.8	22.0	73.8
+ENT [15]	3.9	5.6	39.6	110.7	29.4	21.3	72.5
+LIBRA	**1.5**	**4.5**	37.7	110.1	29.6	22.0	74.6
Transformer [42]	5.4	6.9	35.0	101.5	28.9	21.1	75.3
+ENT [15]	4.4	6.8	38.6	107.1	28.9	20.8	72.9
+LIBRA	**2.3**	**5.0**	33.9	98.7	28.6	20.9	75.7
OSCAR [5]	2.4	3.0	39.4	119.8	32.1	24.0	75.8
+ENT [15]	5.7	**2.8**	41.4	113.0	30.2	21.9	72.8
+LIBRA	**0.3**	4.6	37.2	113.1	31.1	23.2	75.7
ClipCap [94]	1.1	5.6	34.8	103.7	29.6	21.5	76.6
+ENT [15]	3.6	5.1	37.4	101.7	28.4	20.1	73.0
+LIBRA	**-1.5**	**4.5**	33.8	100.6	29.3	21.4	76.0
GRIT [11]	3.1	3.5	42.9	123.3	31.5	23.4	76.2
+ENT [15]	5.2	**3.7**	42.8	120.3	30.8	22.7	74.0
+LIBRA	**0.7**	4.1	40.5	116.8	30.6	22.6	75.9

Table 3.9: Comparison with DCG without masking input captions. Bold numbers denote the best scores in the DCG with/without masking.

	Gender bias ↓		Accuracy ↑	
Model	LIC	Error	SPICE	CLIPScore
NIC [2]	0.5	23.6	13.4	65.2
+DCG w/o mask	**-2.1**	5.9	15.7	70.5
+LIBRA	-0.3	**5.7**	16.5	71.7
SAT [27]	-0.3	9.1	19.2	72.1
+DCG w/o mask	-1.3	4.0	19.8	72.8
+LIBRA	**-1.4**	**3.9**	20.0	73.6
FC [54]	2.9	10.3	18.3	70.0
+DCG w/o mask	0.5	4.4	19.6	72.0
+LIBRA	**-0.2**	**4.3**	19.7	72.9
Att2in [54]	1.1	5.4	20.2	72.6
+DCG w/o mask	0.7	**4.6**	20.6	73.0
+LIBRA	**-0.3**	**4.6**	20.6	73.8
UpDn [4]	4.7	5.6	22.0	73.8
+DCG w/o mask	1.9	4.8	21.9	73.8
+LIBRA	**1.5**	**4.5**	22.0	74.6
Transformer [42]	5.4	6.9	21.1	75.3
+DCG w/o mask	4.4	5.6	20.9	74.9
+LIBRA	**2.3**	**5.0**	20.9	75.7
OSCAR [5]	2.4	3.0	24.0	75.8
+DCG w/o mask	1.9	4.7	23.4	75.8
+LIBRA	**0.3**	**4.6**	23.2	75.7
ClipCap [94]	1.1	5.6	21.5	76.6
+DCG w/o mask	0.5	4.7	21.4	76.2
+LIBRA	**-1.5**	**4.5**	21.4	76.0
GRIT [11]	3.1	3.5	23.4	76.2
+DCG w/o mask	1.8	4.3	22.8	75.3
+LIBRA	**0.7**	**4.1**	22.6	75.9

Table 3.10: Comparison of data used for training DCG. Bold numbers denote the best scores among the types of synthetic datasets.

	Synthesis method			Gender bias ↓	
Model	Swap	T5	Merged	LIC	Error
NIC [2]	-	-	-	0.5	23.6
+LIBRA	✓	✓	-	-0.1	7.5
+LIBRA	-	✓	✓	**-0.3**	**5.7**
+LIBRA	✓	✓	✓	-0.2	6.2
SAT [27]	-	-	-	-0.3	9.1
+LIBRA	✓	✓	-	-2.0	6.2
+LIBRA	-	✓	✓	-1.4	**3.9**
+LIBRA	✓	✓	✓	**-2.3**	4.8
FC [54]	-	-	-	2.9	10.3
+LIBRA	✓	✓	-	0.5	6.5
+LIBRA	-	✓	✓	-0.2	**4.3**
+LIBRA	✓	✓	✓	**-0.9**	5.0
Att2in [54]	-	-	-	1.1	5.4
+LIBRA	✓	✓	-	2.0	6.7
+LIBRA	-	✓	✓	-0.3	**4.6**
+LIBRA	✓	✓	✓	**-1.2**	5.5
UpDn [4]	-	-	-	4.7	5.6
+LIBRA	✓	✓	-	2.3	6.2
+LIBRA	-	✓	✓	1.5	**4.5**
+LIBRA	✓	✓	✓	**1.1**	5.2
Transformer [42]	-	-	-	5.4	6.9
+LIBRA	✓	✓	-	**1.5**	6.9
+LIBRA	-	✓	✓	2.3	**5.0**
+LIBRA	✓	✓	✓	2.6	5.8
OSCAR [5]	-	-	-	2.4	3.0
+LIBRA	✓	✓	-	**-0.8**	6.8
+LIBRA	-	✓	✓	0.3	**4.6**
+LIBRA	✓	✓	✓	0	5.0
ClipCap [94]	-	-	-	1.1	5.6
+LIBRA	✓	✓	-	-1.3	6.8
+LIBRA	-	✓	✓	-1.5	**4.5**
+LIBRA	✓	✓	✓	**-1.7**	5.3
GRIT [11]	-	-	-	3.1	3.5
+LIBRA	✓	✓	-	**-0.8**	6.3
+LIBRA	-	✓	✓	0.7	**4.1**
+LIBRA	✓	✓	✓	0	4.8

Table 3.11: Comparison with random perturbation. Rand. pert. denotes DCG trained on data with random perturbation. Bold numbers denote the best scores in the DCG trained on either biased captions from BCS or captions with random perturbation.

	Gender bias ↓		Accuracy ↑	
Model	LIC	Error	SPICE	CLIPScore
NIC [2]	0.5	23.6	13.4	65.2
+Rand. pert. mask	0.7	7.7	16.4	71.5
+LIBRA	**-0.3**	**5.7**	16.5	71.7
SAT [27]	-0.3	9.1	19.2	72.1
+Rand. pert.	**-1.5**	6.5	19.9	73.4
+LIBRA	-1.4	**3.9**	20.0	73.6
FC [54]	2.9	10.3	18.3	70.0
+Rand. pert.	0.2	6.6	19.8	72.7
+LIBRA	**-0.2**	**4.3**	19.7	72.9
Att2in [54]	1.1	5.4	20.2	72.6
+Rand. pert.	**-0.8**	5.9	20.4	73.7
+LIBRA	-0.3	**4.6**	20.6	73.8
UpDn [4]	4.7	5.6	22.0	73.8
+Rand. pert.	2.2	5.9	21.8	74.4
+LIBRA	**1.5**	**4.5**	22.0	74.6
Transformer [42]	5.4	6.9	21.1	75.3
+Rand. pert.	3.6	6.2	20.7	75.4
+LIBRA	**2.3**	**5.0**	20.9	75.7
OSCAR [5]	2.4	3.0	24.0	75.8
+Rand. pert.	2.0	5.6	22.9	75.4
+LIBRA	**0.3**	**4.6**	23.2	75.7
ClipCap [94]	1.1	5.6	21.5	76.6
+Rand. pert.	0.5	5.9	21.2	75.8
+LIBRA	**-1.5**	**4.5**	21.4	76.0
GRIT [11]	3.1	3.5	23.4	76.2
+Rand. pert.	1.8	5.6	22.4	75.8
+LIBRA	**0.7**	**4.1**	22.6	75.9

Chapter 4

Mitigating Societal Bias Beyond Single Attributes

4.1 Overview

Models trained on biased data can develop prediction rules based on spurious correlations (i.e., associations devoid of causal relationships), perpetuating and amplifying harmful stereotypes [97]. For example, image captioning models may generate gendered captions by associating gender with depicted activities [98], location [99], or objects [29]. Dataset-level bias mitigation aims to reduce spurious correlations between labeled image attributes (e.g., `teddy bear`) and protected groups (e.g., `woman`). Resampling approaches balance the co-occurrence of each attribute with each group [12, 100]. However, models can still exploit correlations between groups and sets of attributes (e.g., `man` with $\{$`dog`, `pizza`, `couch`$\}$), even when individual attributes are balanced [98]. Moreover, spurious correlations extend to unlabeled attributes, which current strategies do not address—e.g., gender disparities in image color statistics [101] or the person-to-object spatial distances [31].

While equal group distributions in real-world datasets are challenging to achieve, generative text-to-image models now enable targeted image modifications [102–104]. For example, bias

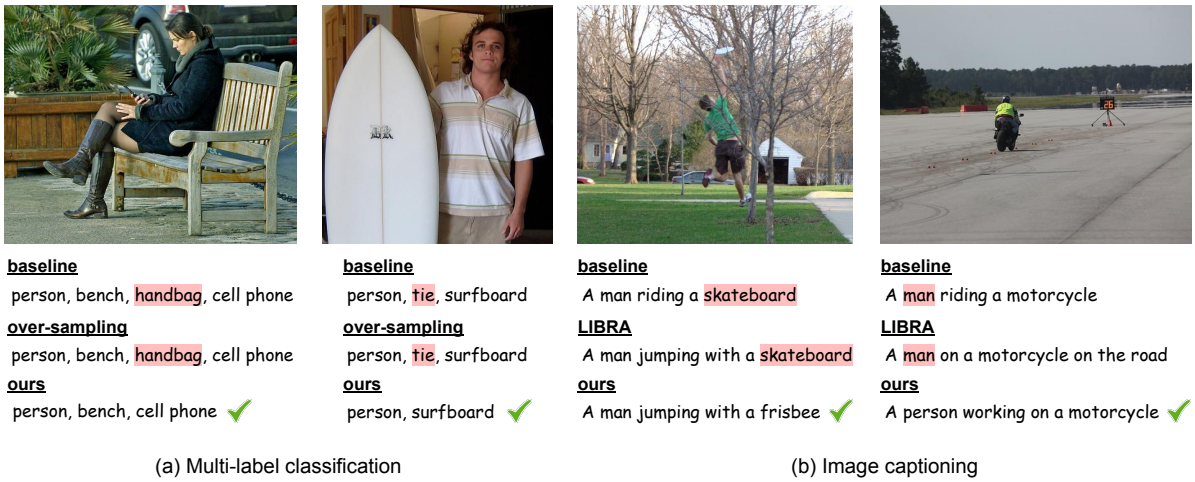


Figure 4.1: (a) Predicted objects by baseline ResNet-50 and with bias mitigation, i.e., over-sampling [12] versus our method. (b) Generated captions by baseline ClipCap and with bias mitigation, i.e., LIBRA [13] versus our method. Incorrect predictions, possibly affected by gender-object correlations, are in red.

detection methods alter image subjects' appearance to assess counterfactual fairness [105] or model bias [106, 107]. However, manipulating individuals' appearances without consent raises significant ethical and privacy concerns [108–113].

To address these challenges, we create training datasets with text-guided inpainting [102], ensuring attribute distributions are independent of protected groups. Using masked person images and text prompts, we generate counterfactual images by inpainting only the masked regions, addressing ethical concerns of altering nonconsensual persons and ensuring equal representation of protected groups across attributes. We introduce data filters to mitigate biases from generative text-guided inpainting models [114–117], evaluating images based on adherence to prompts, preservation of attributes and semantics, and color fidelity, validated by human evaluators. Unlike prior work [12, 39, 98, 100], training on our counterfactual data decorrelates both labeled and unlabeled attributes from protected groups without impacting model performance. Comprehensive evaluations show our approach significantly reduces prediction rules based on spurious correlations in multi-label classification and image captioning across various

architectures (e.g., ResNet-50 [118], Swin Transformer [119]), datasets (COCO [28], OpenImages [120]), and protected groups (gender, skin tone). Our key contributions are summarized as follows:

- Introducing a framework for generating synthetic training datasets with group-independent image attribute distributions.
- Proposing data filtering to mitigate biases introduced by generative inpainting models.
- Conducting quantitative experiments, demonstrating significant bias reduction in classification and captioning tasks compared to baselines.
- Identifying limitations of training on combined real and synthetic datasets, emphasizing the need for cautious synthetic data augmentation.

4.1.1 Related Work

Societal bias in datasets, characterized by demographic imbalances leading to spurious correlations, has been extensively studied [31, 43, 44, 101, 121, 122]. These biases persist and can be exacerbated by multi-label classifiers [34, 39, 97] and image captioning models [6, 75, 99], disproportionately impacting historically marginalized groups such as women and individuals with darker skin tones [77, 123].

Two common approaches to bias mitigation are dataset-level and model-level. Dataset-level approaches leverage generative adversarial networks (GANs), counterfactual training dataset augmentation, and resampling. GANs create synthetic images to balance datasets and mitigate spurious correlations [124–126], counterfactual data augmentation generates alternative scenarios to address biases [127, 128], and resampling balances the co-occurrence of attributes and protected groups [12, 100]. Model-level approaches reduce bias through corpus-level constraints [97], adversarial debiasing [36, 39, 45, 99], domain discriminative/independent training [12], modified loss functions [129–131], and model output editing [13]. However, despite these advancements, existing mitigation methods focus on single labeled attributes, which can

inadvertently increase models' reliance on spurious correlations between protected groups and combinations of attributes [98] or unlabeled attributes [101].

Recent progress in text-to-image generative models has enabled targeted image manipulation [102–104], which can help address bias in multi-modal datasets. Nonetheless, these models have also been shown to perpetuate harmful stereotypes [132–139]. In contrast to prior bias mitigation work, we use text-guided inpainting to generate synthetic training datasets that ensure equal representation of protected groups across all attribute combinations, whether labeled or unlabeled. To mitigate inpainting biases, we propose data filters, producing higher quality and less biased synthetic data. We go beyond previous work focused solely on gender bias mitigation [105–107] by also addressing skin tone biases.

4.2 Method

We create training datasets with group-independent image attribute distributions by using masked person images and text prompts with an off-the-shelf diffusion model, as outlined in Figure 4.2.

4.2.1 Resampled Datasets Are Not Enough

We denote an image by $x \in \mathcal{X}$, a protected group by $g \in \mathcal{G}$, and an image attribute by $a \in \mathcal{A}$. A spurious correlation exists if $p_{\mathcal{X}}(a \mid g) \neq p_{\mathcal{X}}(a)$, indicating biases in the data. Resampling aims to remove these biases by adjusting the sampling process so that $p_{\mathcal{X}}(a \mid g) = p_{\mathcal{X}}(a)$ for all g [12, 34]. This is done using a limited set of labeled attributes $\mathcal{O} \subset \mathcal{A}$, where attributes a are drawn from a distribution $q(a)$ over $\mathcal{O}$ and groups g are drawn from a uniform distribution $u(g)$ over $\mathcal{G}$ such that $\mathcal{X}' = \{x \sim p_{\mathcal{X}}(x \mid g, a) \mid a \sim q(a), g \sim u(g)\}$. This ensures $p_{\mathcal{X}'}(a \mid g) = q(a)$ for $a \in \mathcal{O}$ and $g \in \mathcal{G}$. However, this method has a limitation: it does not account for a being an unlabeled attribute or a combination of labeled and unlabeled attributes, making it difficult to sample x from $p_{\mathcal{X}}(x \mid g, a)$ due to insufficient information about a. In short, while resampling can reduce biases, it is not always enough, especially when dealing with unlabeled or mixed attributes.

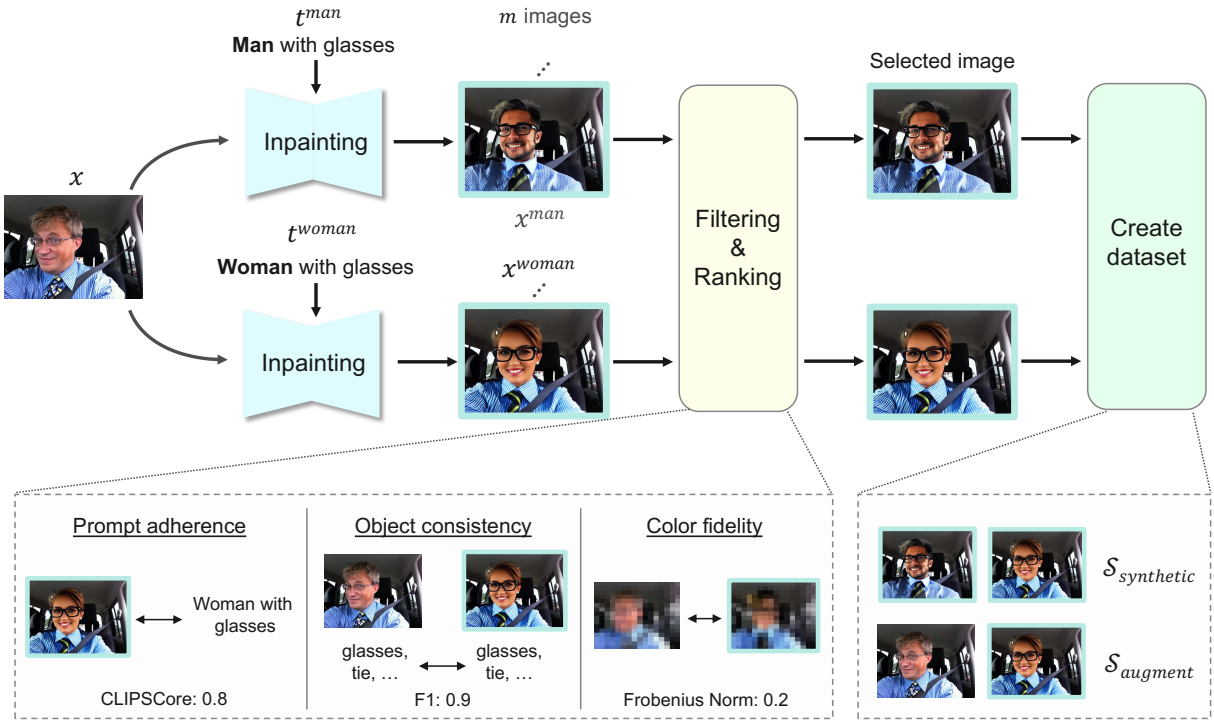


Figure 4.2: Overview of our pipeline for binary gender as a protected attribute. Original images are inpainted to synthesize diverse groups, maintaining consistent context. Synthesized images (highlighted in blue) are ranked using filters to select high-quality, unbiased samples (Module: Filtering & Ranking). Selected images are then used to construct datasets with group-independent image attribute distributions (Module: Create dataset).

4.2.2 Text-Guided Inpainting

Suppose $\mathcal{D} = \{(x_i, \omega_i, a_i, t_i^{(g)}) \mid 1 \leq i \leq n\}$ is a training set, where $x \in \mathbb{R}^d$ is an image, $\omega \in [0,1]^d$ is a person mask, a is a labeled image attribute, a combination of labeled attributes, or an unlabeled attribute, and $t^{(g)}$ is a text prompt containing a protected group-specific word g. To create a dataset with group-independent image attribute distributions, we utilize a text-guided inpainting model [102]. This model, guided by $t^{(g)}$, inpaints ω in x with a synthetic person from protected group g described in $t^{(g)}$. For each tuple in $\mathcal{D}$, we generate $m \in \mathbb{N}^+$

versions for each $g \in \mathcal{G}$, resulting in $m \cdot |\mathcal{G}|$ samples:

$$\mathcal{D}_{\text{synthetic}} = \{(x_i^{(j,g')}, \omega_i, a_i, t_i^{(g')}) \mid 1 \leq i \leq n, g' \in \mathcal{G}, 1 \leq j \leq m\}, \tag{4.1}$$

where $x_i^{(j,g')}$ denotes the j-th inpainted version of $x_i \in \mathcal{X}$ for g' and $t_i^{(g')}$ the modified text prompt where g in $t_i^{(g)}$ is replaced with g'.

4.2.3 Societal Bias Data Filtering

Text-to-image generative models often perpetuate societal biases, portraying certain groups stereotypically, such as depicting women in brighter clothing [114–117]. Since these biases remain largely unaddressed [106, 107], we set $m > 1$ in Equation 4.1 to generate multiple variations for each group. We propose filters to select the least biased inpainted images, evaluating images based on adherence to text prompts, preservation of attributes and semantics, and color fidelity. Specifically, for each tuple (i, g'), we select the highest quality and least biased version among the m versions to create a training dataset:

$$\mathcal{S}_{\text{synthetic}} = \{(x_i^{(j^\star,g')}, \omega_i, a_i, t_i^{(g')}) \in \mathcal{D}_{\text{synthetic}} \mid \forall (i, g'), j^\star\}, \tag{4.2}$$

where $j^\star = \arg\min_j \sum_k c_k \cdot r(s_k^{(i,j,g')})$, $c_k \in \mathbb{R}$ are weights assigned to filters s_k, $s_k^{(i,j,g')}$ is the score obtained from applying filter s_k to image $x_i^{(j,g')}$ for group g', and $r(s_k^{(i,j,g')})$ is the rank of the score for (i, g') in descending order, with lower ranks indicating less bias. Here, $x_i^{(j^\star,g')}$ is the selected inpainted image for tuple (i, g') that minimizes the sum of the ranks of the weighted filter scores, with $j^\star$ representing the index of the selected candidate image for tuple (i, g').

Rather than creating an entire dataset of synthetic samples, we can augment $\mathcal{D}$:

$$\mathcal{S}_{\text{augment}} = \mathcal{D} \cup \{(x_i^{(j^\star,g')}, \omega_i, a_i, t_i^{(g')}) \in \mathcal{D}_{\text{synthetic}} \mid \forall (i, g' \neq g), j^\star\}. \tag{4.3}$$

The condition $g' \neq g$ ensures that we only add inpainted images to $\mathcal{D}$ for groups different from those originally present in x_i. In contrast to resampling, $\mathcal{S}_{\text{synthetic}}$ and $\mathcal{S}_{\text{augment}}$ ensure $p_{\mathcal{X}'}(a \mid$

$g) = p_{\mathcal{X}}(a)$ for all $g \in \mathcal{G}$ without making assumptions about $\mathcal{A}$. Our proposed filters are introduced below.

Prompt Adherence. To evaluate the semantic alignment between $x_i^{(j,g')}$ and $t_i^{(g')}$, we use CLIPScore [7], which computes the cosine similarity between their CLIP embeddings [14]. Formally,

$$s_{\text{prompt}}^{(i,j,g')} = \phi(x_i^{(j,g')}) \cdot \psi(t_i^{(g')}) \in [-1, 1], \tag{4.4}$$

where ϕ and ψ are CLIP's vision and text encoders, respectively. If $s_{\text{prompt}}^{(i,j,g')} > s_{\text{prompt}}^{(i,j',g')}$, then $x_i^{(j,g')}$ better reflects the content described in $t_i^{(g')}$.

Object Consistency. To prevent the introduction of spurious correlations, such as generating objects not mentioned in $t_i^{(g')}$ or reinforcing stereotypes [114–116], we assess the object similarity between predicted objects in $x_i^{(j,g')}$ and x_i. Concretely, we compute the F1 score [140] using a pretrained object detector [141], denoted η:

$$s_{\text{object}}^{(i,j,g')} = \text{F1}[\eta(x_i^{(j,g')}), \eta(x_i)] \in [0, 1]. \tag{4.5}$$

If $s_{\text{object}}^{(i,j,g')} > s_{\text{object}}^{(i,j',g')}$, then $x_i^{(j,g')}$ better preserves the integrity of the original unmasked scene in x_i.

Color Fidelity. Generative models can introduce subtler biases [114, 116], including those related to color [101]. Addressing color biases is crucial as color choices can implicitly carry cultural or gendered connotations. To mitigate this, we downsample $x_i^{(j,g')}$ and x_i to 14×14 pixels to focus on color rather than fine details, then measure the color difference using the Frobenius norm:

$$s_{\text{color}}^{(i,j,g')} = \|(x_i^{(j,g')})_{\downarrow 14\times 14} - (x_i)_{\downarrow 14\times 14}\|_{\text{F}}^{-1}. \tag{4.6}$$

If $s_{\text{color}}^{(i,j,g')} > s_{\text{color}}^{(i,j',g')}$, then $x_i^{(j,g')}$ has better color fidelity to the original unmasked scene in x_i.

	ResNet-50			Swin-T			ConvNeXt-B		
	mAP	Ratio	Leakage	mAP	Ratio	Leakage	mAP	Ratio	Leakage
Original	<u>66.4</u>	6.3	13.4	**72.8**	4.0	14.3	**76.3**	4.6	18.2
Adversarial	63.3	—	**3.3**	67.8	—	**4.4**	69.6	—	**4.7**
DomDisc	57.4	4.1	15.4	65.4	4.6	16.8	68.8	4.5	19.1
DomInd	60.4	2.8	10.4	67.9	3.8	11.4	72.6	5.9	15.0
Upweight	64.9	9.1	8.3	71.5	6.3	9.8	75.0	5.6	12.9
Focal	66.1	6.3	12.0	<u>72.2</u>	3.8	13.3	<u>76.2</u>	3.8	16.2
CB	63.0	4.3	10.9	69.6	3.5	12.3	73.8	3.5	14.7
GroupDRO	64.1	3.0	11.4	70.8	<u>1.5</u>	12.6	75.3	4.2	16.4
Over-sampling	62.6	3.8	9.7	69.9	2.6	10.5	73.5	3.4	13.7
Sub-sampling	58.3	<u>2.0</u>	12.2	64.4	1.8	11.6	66.3	<u>2.2</u>	18.2
$\mathcal{S}_{\text{augment}}$ (Ours)	**66.9**	4.6	8.1	**72.8**	3.1	10.5	**76.3**	<u>2.2</u>	11.3
$\mathcal{S}_{\text{synthetic}}$ (Ours)	66.0	**1.1**	<u>7.5</u>	71.9	**1.4**	<u>8.4</u>	75.5	**1.2**	<u>8.2</u>

Table 4.1: Classification performance and gender bias scores of ResNet-50, Swin-T, and ConvNeXt-B backbones on COCO. Ratio is inapplicable to Adversarial due to its gender prediction module for mitigation. **Bold** and <u>underline</u> represent the best and second-best, respectively. For an unbiased model, Ratio $= 1$ and Leakage $= 0$.

4.3 Experiments

Building on prior research [6, 39, 45, 97–99, 142], we evaluate our synthetic dataset creation method on multi-label image classification and image captioning tasks using quantitative metrics, human studies, qualitative comparisons, and effectiveness analysis. Evaluations are conducted on test sets of real data.

Implementation Details. We inpaint the largest person in the image based on bounding box size, and if the second largest person exceeds 55,000 pixels, we also inpaint that region, using the `person` label for COCO. For image generation, we create $m = 30$ inpainted images per

	ClipCap				BLIP-2				Transformer			
	M	CS	Ratio	LIC	M	CS	Ratio	LIC	M	CS	Ratio	LIC
Original	**29.1**	<u>75.1</u>	2.5	2.2	**29.5**	75.1	5.7	4.7	<u>26.9</u>	<u>71.5</u>	4.7	4.7
LIBRA	28.9	74.9	6.5	<u>0.5</u>	29.0	**75.4**	6.3	**1.9**	**27.4**	**73.4**	6.7	2.3
Over-sampling	28.6	74.7	3.2	3.5	28.7	74.1	3.8	3.0	26.2	70.6	4.1	1.6
Sub-sampling	28.0	74.0	<u>1.4</u>	4.1	28.3	74.5	<u>1.4</u>	3.2	25.0	69.7	<u>2.0</u>	3.9
$\mathcal{S}_{\text{augment}}$ (Ours)	<u>29.0</u>	75.0	2.5	1.7	<u>29.4</u>	<u>75.3</u>	2.9	3.8	26.2	71.1	2.6	<u>1.5</u>
$\mathcal{S}_{\text{synthetic}}$ (Ours)	28.5	**75.3**	**1.3**	**0.3**	29.3	75.0	**1.2**	<u>2.5</u>	25.7	70.9	**1.4**	**0.5**

Table 4.2: Captioning quality and gender bias scores of ClipCap, BLIP-2, and Transformer backbones on COCO. M and CS denote METEOR and CLIPScore. **Bold** and <u>underline</u> represent the best and second-best, respectively. For an unbiased model, Ratio $= 1$ and LIC $= 0$.

group (e.g., `woman`, `man`) using guidance scales of 7.5, 9.5, and 15.0 to ensure diversity. Filter weights are set to 1 (i.e., $c_k = 1$ for all k), contributing equally. Results are based on five models trained with different random seeds.

4.3.1 Multi-Label Classification

Experimental Setup. We focus on gender bias using the COCO dataset, retaining only images with gender-specific terms (e.g., `woman`, `man`) in their captions. This results in 28,487/13,487 train/test samples. We focus on objects co-occurring with these terms, yielding 51 objects. ResNet50, Swin Transformer Tiny (Swin-T), and ConvNext models are fine-tuned using early stopping. Performance is assessed using mean average precision (mAP). Bias is quantified using leakage and ratio. Leakage measures how much the model's predictions amplify the group's information compared to the ground truth. A gender classifier $f_g(y)$, predicting gender group g from input y (i.e., set of objects), is trained on a training set $\mathcal{T} = \{(y, g)\}$. For the test set $\mathcal{T}'$, the model's leakage score is:

$$\text{LK}_{\text{M}} = \frac{1}{|\mathcal{T}'|} \sum_{(y,g)\in\mathcal{T}'} f_g(y) \mathbb{1}\left[\arg\max_{g'} f_{g'}(y) = g\right] \tag{4.7}$$

The leakage score for the original dataset, $\mathrm{LK_D}$, is similarly computed. The final leakage is Leakage $= \mathrm{LK_M} - \mathrm{LK_D}$. Higher leakage indicates greater model exploitation of protected group information. Ratio measures the exploitation of attribute information for group prediction. By masking individuals in test images and measuring the bias in group predictions (e.g., #man-to-#woman ratio), deviations from a ratio of 1 indicate attribute exploitation. We report Ratio $= \max(r, r^{-1})$, where r is the observed ratio. This captures the magnitude of deviation from unbiased predictions consistently.

We compare our method with existing bias mitigation techniques, including dataset-level methods (Over-sampling [12], Sub-sampling [100]) and model-level methods such as adversarial debiasing [39] (Adversarial), domain-independent training [12] (DomInd), domain discriminative training [12] (DomDisc), loss upweighting [143] (Upweight), focal loss [129] (Focal), class-balanced loss [130] (CB), and group DRO [131] (GroupDRO). Additional results on the OpenImages dataset and skin tone bias mitigation are provided in Section 4.6.1, demonstrating consistent conclusions.

Results. Results are shown in Table 4.1. Our method, $\mathcal{S}_{\text{synthetic}}$, achieves the best balance by significantly improving both ratio and leakage while maintaining a high mAP. Specifically, $\mathcal{S}_{\text{synthetic}}$ achieves a near-ideal ratio of 1.1, low leakage of 7.5, and an mAP of 66.0 for ResNet-50, with similar trends observed for Swin-T and ConvNeXt-B.

Adversarial debiasing achieves lower leakage scores by removing gender information from intermediate representations. However, this method reduces mAP, indicating that object information may also be inadvertently removed. Over-sampling and sub-sampling methods address class imbalance but at the cost of model performance. Sub-sampling, in particular, reduces the ratio compared to over-sampling but results in worse mAP and increased leakage. This is likely due to the loss of diversity and information in the training data, which forces the model to rely more on the remaining features, increasing the influence of protected attributes.

In contrast, $\mathcal{S}_{\text{synthetic}}$ generates diverse, high-quality synthetic samples, effectively balancing bias and variance. This approach avoids the pitfalls of other methods, resulting in superior

performance metrics. While $\mathcal{S}_{\text{augment}}$ performs similarly to the original dataset, it performs worse in terms of ratio and leakage compared to $\mathcal{S}_{\text{synthetic}}$.

4.3.2 Image Captioning

Experimental Setup. Using the COCO dataset (Table 4.3.1), we benchmark captioning models ClipCap, BLIP-2, and Transformer, which are fine-tuned using early stopping. Performance is evaluated with METEOR and CLIPScore. Bias is quantified using LIC and ratio, where LIC is a leakage-based metric that assesses the generation of group-stereotypical captions compared to ground-truth captions (i.e., y is a caption in Equation 4.7), and predicted group-related terms (e.g., `woman`) in captions used to compute ratio.

Bias mitigation baselines include dataset-level methods (Over-sampling, Sub-sampling) and the current state-of-the-art model-level method LIBRA [13]. LIBRA is a model-agnostic debiasing framework designed to mitigate bias amplification in image captioning by synthesizing gender-biased captions and training a debiasing caption generator to recover the original captions. Detailed results for skin tone bias mitigation, along with fine-tuning specifics, are provided in Section 4.6.2, showcasing the generalizability of our approach.

Results. Results are shown in Table 4.2. Our method, $\mathcal{S}_{\text{synthetic}}$, significantly improves both ratio and LIC while maintaining high METEOR and CLIPScore values. Specifically, $\mathcal{S}_{\text{synthetic}}$ achieves a near-ideal ratio of 1.3, low LIC of 1.2, and a METEOR score of 29.3 for BLIP-2, with similar trends observed for ClipCap and Transformer.

While LIBRA effectively reduces LIC, it shows an increase in the ratio metric, indicating a trade-off between debiasing effectiveness and caption quality. Over-sampling and sub-sampling methods result in varying degrees of performance. Sub-sampling showed improved bias metrics compared to over-sampling but results in worse METEOR scores, especially for the Transformer model.

As in the multi-label classification task, we observe that although $\mathcal{S}_{\text{augment}}$ significantly reduces bias compared to using the original dataset, there is a significant gap between it and

	ResNet-50			Swin-T			ClipCap			BLIP-2		
	$\text{Ratio}_{\text{orig}}$	$\text{Ratio}_{\text{inp}}$	Δ	$\text{Ratio}_{\text{orig}}$	$\text{Ratio}_{\text{inp}}$	Δ	$\text{Ratio}_{\text{orig}}$	$\text{Ratio}_{\text{inp}}$	Δ	$\text{Ratio}_{\text{orig}}$	$\text{Ratio}_{\text{inp}}$	Δ
Original	3.5	3.0	14.3	3.1	2.6	16.1	2.3	2.5	8.7	2.3	2.4	4.4
$\mathcal{S}_{\text{augment}}$	3.7	1.5	59.5	3.2	0.6	81.3	2.5	0.8	68.0	2.3	1.8	21.7
$\mathcal{S}_{\text{synthetic}}$	1.9	1.8	5.3	2.1	2.0	4.8	1.7	1.6	5.9	1.8	1.7	5.6

Table 4.3: Comparison of the original ($\text{Ratio}_{\text{orig}}$) and inpainted ($\text{Ratio}_{\text{inp}}$) versions of the COCO test set. The relative difference is denoted by $\Delta = 100 \cdot |\frac{\text{Ratio}_{\text{orig}} - \text{Ratio}_{\text{inp}}}{\text{Ratio}_{\text{orig}}}|\%$. A larger Δ signifies a greater change.

$\mathcal{S}_{\text{synthetic}}$ in terms of bias mitigation.

4.3.3 Analysis of Synthetic Artifacts

Recent studies show that text-to-image models introduce synthetic artifacts in images, which models may exploit [144–146]. Our observations in Sections 4.3.1 and 4.3.2 suggest that bias persists with $\mathcal{S}_{\text{augment}}$, which augments the dataset with counterfactual images to balance group distributions. We hypothesize that $\mathcal{S}_{\text{augment}}$ may lead to shortcut learning due to spurious correlations between minoritized groups and inpainted artifacts. In contrast, $\mathcal{S}_{\text{synthetic}}$ distributes artifacts equally across all groups, avoiding this issue.

To test this, we create a test set by inpainting random body parts using COCO-WholeBody annotations [147]. Given an image, its caption, and body part annotations (e.g., left hand, right hand, head), we randomly select a body part, create a mask using the Segment Anything Model [148], and perform inpainting with the caption as a prompt. We evaluate the consistency of ratios between the original and synthetic test sets; a gap indicates the exploitation of synthetic artifacts for gender prediction.

Table 4.3 presents scores for multi-label classification (ResNet-50, Swin-T) and image captioning (ClipCap, BLIP-2). The table includes the ratio of gender predictions (#`man`-to-#`woman`) for the original test set ($\text{Ratio}_{\text{orig}}$) and the inpainted test set ($\text{Ratio}_{\text{inp}}$), along with the relative difference (Δ) between these ratios. Results show a significant shift in gender predictions with $\mathcal{S}_{\text{augment}}$-trained models. Despite identical gender ratios in the original and inpainted

	Object	Color	Skin	Gender	CS
$s_{\text{prompt}} + s_{\text{object}} + s_{\text{color}}$	**0.57**	0.46	<u>0.29</u>	0.95	**75.3**
$s_{\text{prompt}} + s_{\text{object}}$	0.49	0.50	0.20	**0.99**	74.8
$s_{\text{prompt}} + s_{\text{color}}$	0.45	**0.56**	0.21	0.94	<u>75.2</u>
$s_{\text{object}} + s_{\text{color}}$	<u>0.53</u>	<u>0.52</u>	0.20	0.96	74.8
s_{prompt}	0.32	0.46	0.26	<u>0.97</u>	75.1
s_{object}	0.36	0.43	0.25	0.95	74.5
s_{color}	0.52	0.50	**0.30**	0.95	74.6
No filter	0.09	0.07	0.18	0.94	74.6

Table 4.4: Human evaluation and captioning quality (CLIPScore, CS in short) for each filter combination. Higher values indicate better alignment with original images. **Bold** and <u>underline</u> represent the best and second-best score for each metric.

test sets (both set at 2.3), models trained with $\mathcal{S}_{\text{augment}}$ predict `woman` much more frequently for the inpainted test set, indicated by the large relative differences. In contrast, models trained solely on synthetic data ($\mathcal{S}_{\text{synthetic}}$) show minimal relative differences, indicating consistent gender predictions across original and inpainted test sets.

Figure 4.3 shows examples of synthetic images and predictions by ClipCap (trained on $\mathcal{S}_{\text{augment}}$ or $\mathcal{S}_{\text{synthetic}}$). The examples demonstrate inconsistent gender predictions with $\mathcal{S}_{\text{augment}}$; specifically, the model tends to predict `woman` for the inpainted test images, evidencing exploitation of synthetic artifacts.

4.3.4 Human Filter Evaluation

We conduct human evaluations on Amazon Mechanical Turk [149] to evaluate the effectiveness of our filters, aiming to determine if our filters prevent additional biases from inpainting models and ensure high-quality images. For 300 randomly selected original images, we analyze inpainted images chosen by each filter combination. Evaluations focus on the similarity of 1) held/nearby objects, 2) object color, and 3) skin tone compared to the original images. Workers assess differences between original and synthetic images for objects and their color, and selected skin tone classes using the Monk Skin Tone Scale [150, 151]. Additionally, workers

verify accurate gender depiction through a sentence gap-filling exercise (e.g., "`A _____ with a dog.`"), where they must choose a protected group term to complete the sentence. More details are in Section 4.6.3.

For the evaluation of the similarity of objects and their colors, scores are computed as the proportion of times the inpainted images are rated as similar. Regarding the skin tone and gender evaluations, the scores are calculated as the proportion of matching responses from workers between the original and inpainted images. All the scores range from 0 to 1.

Table 4.4 summarizes the human evaluation and captioning performance of ClipCap trained on $\mathcal{S}_{\text{synthetic}}$ (CS), with images selected by each filter. Notably, using all filters consistently received higher ratings across most criteria. In contrast, randomly selecting images without any filtering often leads to synthetic images differing significantly from the originals. This indicates that our filters are effective in mitigating additional biases introduced by the inpainting model. Furthermore, CLIPScore shows that using all filters improves captioning performance, highlighting its effectiveness in selecting higher-quality images.

4.3.5 Inherited Biases

To further discuss the potential biases introduced by the models used in our method, we conduct several assessments. First, for the object detector, we run Detic [141] on both real and synthetic images, achieving similar mAP scores of 32.0 for real images and 32.3 for synthetic images, indicating consistent performance. Second, addressing biases in CLIP, we acknowledge the potential biases inherent in the model. However, our use of object- and color-based filters helps mitigate these biases. Additionally, image classification and captioning results verify that our method effectively reduces gender and skin tone biases. Lastly, for the inpainting model, our filters effectively remove synthetic images that deviate from the prompt, alter color statistics, or introduce undescribed objects, as shown in Table 4.4. These assessments confirm that our method successfully mitigates biases without compromising performance.

4.3.6 Qualitative Results

We present qualitative examples of bias mitigation by applying our method ($\mathcal{S}_{\text{synthetic}}$) in Figure 4.1. The results show that training models on $\mathcal{S}_{\text{synthetic}}$ produces less biased outputs. For instance, in the classification task, the baseline ResNet-50 model and the over-sampling model incorrectly predict `tie`, due to its frequent co-occurrence with `man` in the training set. In contrast, $\mathcal{S}_{\text{synthetic}}$ results in a gender bias-free prediction. Image captioning results further validate our approach. The baseline ClipCap model and LIBRA model generate the man-stereotypical word `skateboard`, whereas our method correctly predicts the object `frisbee`.

In Figure 4.4, we also present the best and worst inpainted images for each filter (prompt adherence, object consistency, and color fidelity), as well as their combination (overall). The results demonstrate each filter's effectiveness, and combining them selects a high-quality image that closely resembles the original. For instance, the image judged worst by the object consistency filter lacks the object the `man` is holding, while the color fidelity filter's worst image shows significant color changes in the `man`'s clothing. Combining these filters helps select an inpainted image that minimizes additional bias and closely matches the original.

4.4 Conclusion

We present a dataset-level bias mitigation pipeline that effectively reduces gender and skin tone biases by ensuring group-independent attribute distribution using synthetic-only images. Our findings indicate that mixing real and synthetic images introduces spurious correlations, underscoring the need for caution when augmenting datasets with synthetic data. Our work highlights the potential of synthetic data in bias mitigation and suggests further exploration into optimizing synthetic data generation and integration techniques for increased bias reduction.

Limitations

Binarized Group Classes and Intersectional Bias Analysis. While acknowledging that gender and skin tone exist on a spectrum, our data limitations necessitated a focus on binarized groups (i.e., `man`, `woman`). Our focus on gender and skin tone biases was driven by:

- Prevalence in Literature: Gender and skin tone biases have been extensively investigated in previous works, providing a robust foundation for our study [6, 39, 97, 98].
- Availability of Annotations: Current datasets primarily include annotations for gender and skin tone, limiting our ability to extend to other attributes [6].

However, our method can be extended to handle intersectional attributes (e.g., gender and skin tone) by inpainting with combinations of attributes (e.g., {`woman`, `darker-skinned`}, {`woman`, `lighter-skinned`}, {`man`, `darker-skinned`}, {`man`, `lighter-skinned`}). We leave this extension for future work to ensure a more comprehensive and inclusive analysis of biases.

Risks of Using Pre-trained Models. As discussed in Section 4.3.5, the pre-trained models employed in our framework (e.g., inpainting model, object detector) may introduce inherent biases. While our analysis in Section 4.3.5 confirmed that these models do not adversely affect our method based on our evaluations, it is possible that some biases were not detected. Specifically, we propose the following steps for future work:

- Developing and integrating additional filtering techniques to detect and mitigate subtle biases.
- Exploring the use of less biased models, such as debiased versions of CLIP [68].

Residual Bias. Our experimental results demonstrated that our method significantly mitigates societal bias compared to existing methods. However, bias is not completely eliminated (e.g., leakage is not zero). Future work could explore further debiasing by optimizing the weight of

each filter (currently, all filters are equally weighted), introducing additional filters, and combining our method with existing bias mitigation techniques (e.g., focal loss).

Extending to Additional Protected Groups. Due to a lack of annotations for other protected attributes, our focus in this work is on gender and skin tone biases. Nevertheless, our pipeline is applicable to various protected attributes, such as age (e.g., "`A` **`woman`** `with a dog`" → "`An` **`elderly`** `woman with a dog`"). Future research should explore the application of our method to additional protected attributes.

Ethics Statement

Our research involves the manipulation of image data to mitigate societal bias, raising important ethical considerations. We address these concerns by creating synthetic images that completely inpaint over identifiable individuals, thereby respecting privacy and consent without altering their appearance. Our approach aims to promote fairness and equity by ensuring diverse and unbiased representation in image datasets. We acknowledge the potential biases inherent in the pre-trained models used and have implemented filters to mitigate these biases as much as possible. Future work should continue to explore ethical guidelines and safeguards to ensure the responsible use of generative models in research.

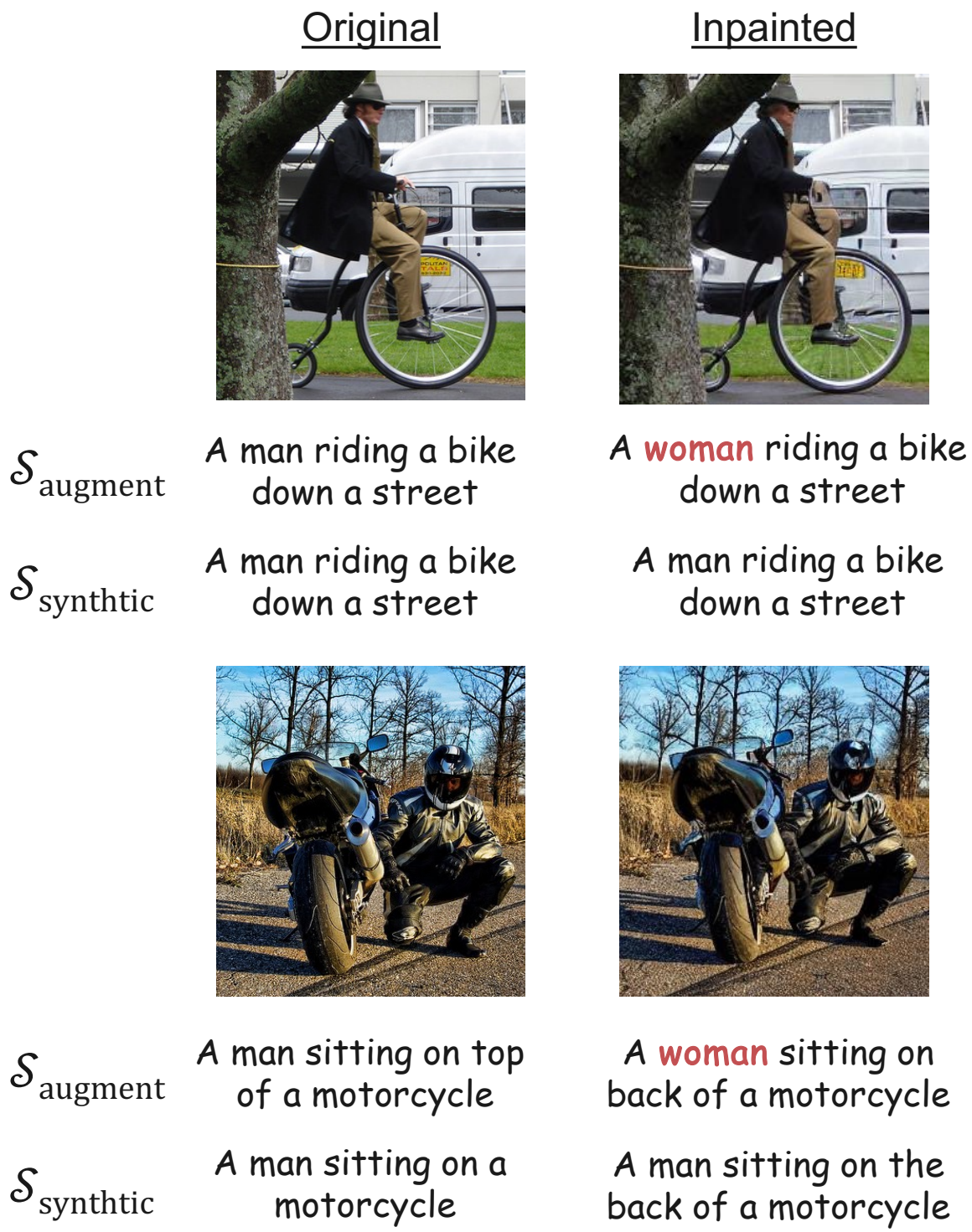


Figure 4.3: Predicted captions for the original (left) and inpainted (right) test images.

Figure 4.4: Best/worst inpainted images for each filter in Section 4.2.3 and their combination (overall).

Appndix

4.5 Method Details

4.5.1 Image Generation Settings

Selection of People for Inpainting. Following the previous works [6, 152], we apply inpainting to a person with the largest bounding box. In addition, if the second largest person's box is larger than $55,000$ pixels, the region is also inpainted. For COCO, we do this by using the `person` label and corresponding bounding boxes. For OpenImages, we use person-bounding boxes presented in More Inclusive Annotations for People (MIAP) annotations [153], then we generate person masks within the boxes using Segment Anything Model [148].

Parameters of Image Generation. In Section 4.2.2, we generate $m = 30$ inpainted images for each group (e.g., $\{\texttt{woman}, \texttt{man}\}$ for binary gender). When generating the images, we use three different guidance scale parameters (7.5, 9.5, and 15.0) to generate diverse inpainted images (i.e., generating 10 images for each guidance scale). We use 6 NVIDIA A100-PCIE-40GB GPUs, resulting in a total of 72 hours to finish synthesizing images.

4.5.2 Visual examples of inpainted images & failure cases

We show the visual examples of the inpainted images after filtering in Figure 4.5 (for binary gender) and Figure 4.6 (for binary skin tone). The examples show that the inpainted images depict the target groups (e.g., `woman` and `darker-skinned`), keeping the rest fixed. In some cases, artifacts are noticeable, which enables us to identify synthetic images (e.g., the details of the faces are not clear), but they do not affect the downstream performance, as shown in the main paper.

Figure 4.5: Examples of inpainted images for binary gender.

4.6 Experimental Settings and Additional Results

4.6.1 Multi-Label Classification

Datasets. We use COCO [28] and OpenImages [120]. Following previous works [97, 98], we focus on attributes co-occurring with `woman` or `man` more than 100 times and remove person-related classes (e.g., `person` class), resulting in 51 and 126 attributes for COCO and OpenImages, respectively. The list of the attributes is as follows:

COCO: {`sink`, `refrigerator`, `laptop`, `surfboard`, `vase`, `bottle`, `remote`, `donut`, `motorcycle`, `car`, `chair`, `suitcase`, `tv`, `knife`, `fork`, `couch`, `bus`, `toothbrush`, `bicycle`, `tie`, `clock`, `microwave`, `teddy bear`, `frisbee`, `spoon`, `dog`, `truck`, `bench`, `backpack`, `skis`, `horse`, `sandwich`, `bed`, `handbag`, `umbrella`, `pizza`, `book`, `dining table`, `traffic light`, `banana`, `potted plant`, `tennis racket`, `cat`, `sports ball`, `kite`, `cake`, `wine glass`, `bowl`, `cup`, `oven`, `cell phone`}.

OpenImages: {`goggles`, `building`, `cloud`, `smile`, `tree`, `sunglasses`, `light`, `t-shirt`, `glasses`, `water`, `forehead`, `wall`, `sky`, `tire`, `roof`, `road`, `wheel`, `vehicle`,

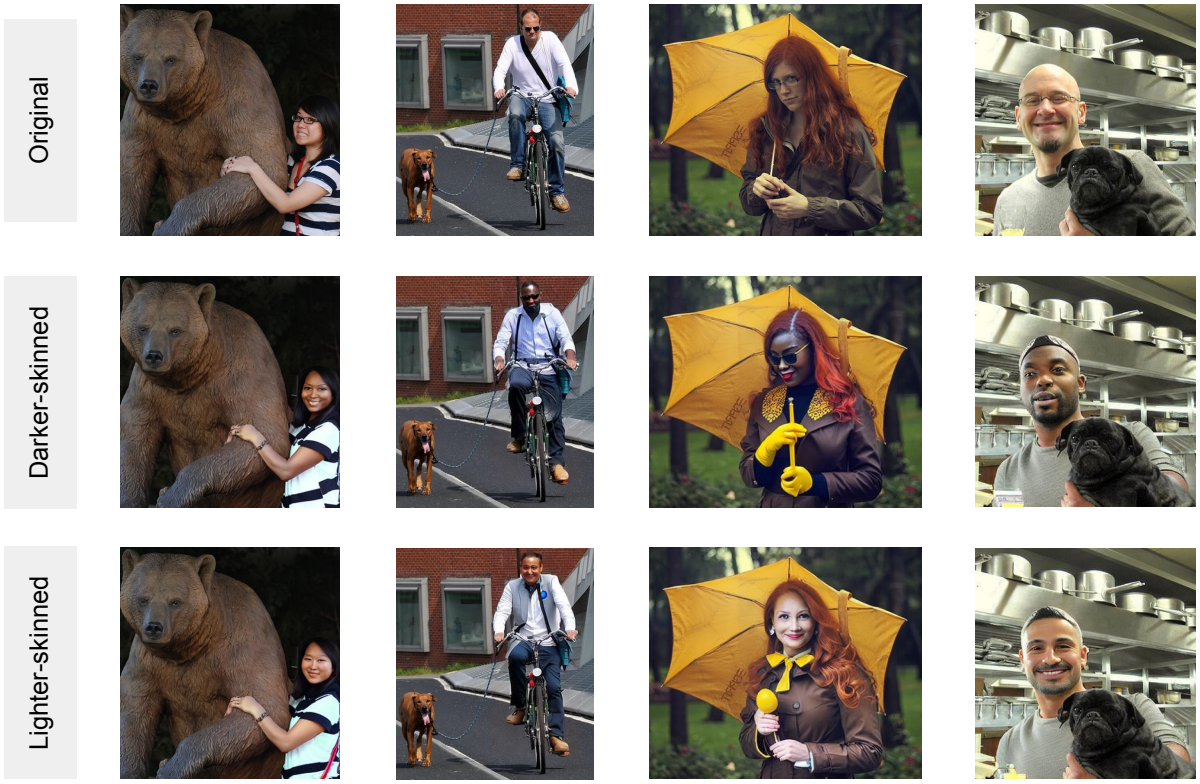


Figure 4.6: Examples of inpainted images for binary skin tone.

land vehicle, car, tie, furniture, microphone, suit, clothing, fence, jeans, trousers, shirt, footwear, flooring, outerwear, coat, ceiling, floor, jacket, table, house, couch, mammal, hat, shoe, sports uniform, baseball (sport), cap, baseball cap, bag, drawing, sun hat, musical instrument, baby, window, door, sweater, lake, chair, tableware, bottle, drink, handwriting, paper, food, tent, concert, drum, guitar, glove, sports equipment, blazer, art, painting, dress, flower, sneakers, screenshot, watercraft, beach, animal, grass family, plant, soil, desk, poster, bus, computer, personal computer, watch, mountain, helmet, bicycle helmet, bicycle wheel, bicycle, curtain, dance, football, ball (object), soccer, wedding dress, jewellery, bride, office building, laptop, toddler, shorts, hiking, fashion accessory, fedora, swimming, swimwear, camera, playground, weapon, ship, statue, boat,

	ResNet-50			Swin-T			ConvNeXt-B		
	mAP	Ratio	Leakage	mAP	Ratio	Leakage	mAP	Ratio	Leakage
Original	<u>42.3</u>	5.2	18.9	<u>45.3</u>	4.3	20.9	<u>46.0</u>	5.0	22.7
Adversarial	37.5	—	**8.3**	40.8	—	**11.3**	40.4	—	**12.3**
DomDisc	40.7	3.7	20.6	43.6	4.6	22.1	42.9	4.1	21.9
DomInd	40.3	3.7	19.1	42.7	3.5	20.2	43.4	2.6	22.0
Upweight	41.3	6.5	<u>13.1</u>	44.7	5.8	17.9	45.3	7.4	18.0
Focal	**43.0**	4.6	18.7	**45.4**	4.4	21.3	45.4	4.0	22.3
CB	40.5	5.2	18.0	42.6	3.9	19.8	43.9	4.6	21.5
GroupDRO	<u>42.3</u>	4.2	18.9	45.1	4.2	20.9	**46.1**	3.4	22.5
Over-sampling	38.5	3.3	15.0	41.1	4.0	<u>16.1</u>	41.7	5.2	18.4
Sub-sampling	38.3	<u>2.2</u>	18.3	41.2	<u>2.1</u>	19.8	39.8	2.8	21.7
$\mathcal{S}_{\text{augment}}$ (Ours)	42.0	1.9	16.0	44.9	2.4	18.0	45.5	2.6	19.0
$\mathcal{S}_{\text{synthetic}}$ (Ours)	41.4	**1.1**	14.6	44.4	**2.0**	17.6	44.7	**1.3**	<u>17.9</u>

Table 4.5: Classification performance and gender bias scores of ResNet-50, Swin-T, and ConvNeXt-B backbones on OpenImages. Ratio is inapplicable to Adversarial due to its gender prediction module for mitigation. **Bold** and <u>underline</u> represent the best and second-best, respectively. For an unbiased model, Ratio $= 1$ and Leakage $= 0$.

`fast food`, `flag`, `soft drink`, `book`, `auto part`, `snow`, `carnivore`, `dog`, `horse`, `motorcycle`, `pole dance`}.

Training. The models (ResNet-50 [118], Swin-T [119], and ConvNeXt-Base [154]) are initialized with ImageNet [155] pre-training, and fine-tuned with early stopping using a validation set split from the training set (20% of the training set). The optimizer is Adam [156], batch size is 32, and a learning rate is 1×10^{-5}. For binary gender, the classification layers predict both protected groups (i.e., {`woman`, `man`}) and object classes. For binary skin tone, the models only predict object classes as ground-truth skin tone labels are not available.

	ResNet-50		Swin-T		ConvNeXt-B	
	mAP	Leakage	mAP	Leakage	mAP	Leakage
Original	**65.8**	3.2	**72.2**	7.1	**75.9**	7.2
$\mathcal{S}_{\text{synthetic}}$ (Ours)	65.2	**2.3**	71.4	**3.7**	74.5	**5.9**

Table 4.6: Classification performance and skin tone bias scores of ResNet-50, Swin-T, and ConvNeXt-B backbones on COCO. **Bold** represents the best. For an unbiased model, Ratio $= 1$ and Leakage $= 0$.

Results for OpenImages. We show the complete results of the experiments in the main paper: gender bias on OpenImages (Table 4.5). The results show that all the insights described in the main paper are consistent across the datasets.

Results for skin tone bias. Previous bias mitigation methods face a significant limitation, requiring protected group labels for all training set samples [39, 97, 100]. They typically focus on gender as a protected attribute due to its prevalence in captions [152], allowing for label inference through gender-related terms. In contrast, $\mathcal{S}_{\text{synthetic}}$ applies to attributes without labels, such as skin tone. We use our pipeline (excluding the color fidelity filter, as we aim to modify skin tone) on binary skin tone categories (i.e., $\mathcal{G} = \{\texttt{darker-skinned}, \texttt{lighter-skinned}\}$) using COCO. We evaluate skin tone bias using *leakage* only since *ratio* requires models to predict protected groups, and there are no skin tone annotations for the COCO training set. Results are shown in Table 4.6, demonstrating consistent conclusions with gender bias.

4.6.2 Image Captioning

Training. We benchmark three captioning models: ClipCap [94], BLIP-2 [157], and Transformer (i.e., the Transformer-based encoder-decoder model composed of Vision Transformer [158] and GPT-2 [92]). As with multi-label classification, we train the models with early stopping. Specifically, for ClipCap, we follow the official implementation regarding the training

	ClipCap			BLIP-2			Transformer		
	M	CS	LIC	M	CS	LIC	M	CS	LIC
Original	**29.4**	75.3	4.6	**27.1**	**73.9**	2.2	**27.0**	**71.5**	5.3
$\mathcal{S}_{\text{synthetic}}$ (Ours)	29.1	**75.4**	**3.7**	26.8	73.6	**2.0**	26.5	71.0	**4.7**

Table 4.7: Captioning quality and skin tone bias scores of ClipCap, BLIP-2, and Transformer backbones on COCO. M and CS denote METEOR and CLIPScore. **Bold** represents the best. For an unbiased model, Ratio $= 1$ and LIC $= 0$.

settings. For BLIP-2 and Transformer, we use the implementation in Hugging Face [159]. We use the AdamW optimizer [160] with a learning rate of $2 \times 10^{-6}/1 \times 10^{-4}$ and batch size of $8/64$ for BLIP-2 and Transformer, respectively.

Results for skin tone. We show the results of the experiments for skin tone bias mitigation in Table 4.7. The results show that the insights in the main paper are mostly consistent across the protected groups.

4.6.3 Human Filter Evaluation

In Figures 4.7, 4.8, and 4.9, we present example tasks for human evaluation conducted on Amazon Mechanical Turk (AMT) [149]. This evaluation assesses how well each combination of filters identifies desirable inpainted images. Figure 4.7 shows the user interface for evaluating the similarity of held/nearby objects and their colors between the original (left) and inpainted (right) images. Figure 4.8 asks workers to select a skin tone class using the Monk Skin Tone Scale [150,151]. We conduct this evaluation on both original and inpainted images and compute the degree of agreement between them. Figure 4.9 verifies if *perceived* gender is accurately depicted—according to the AMT worker—in the inpainted images through gap-filling, where workers must choose a protected group term to complete the sentence. Each assignment pays \$0.07, with a total participant compensation of approximately \$2,000.

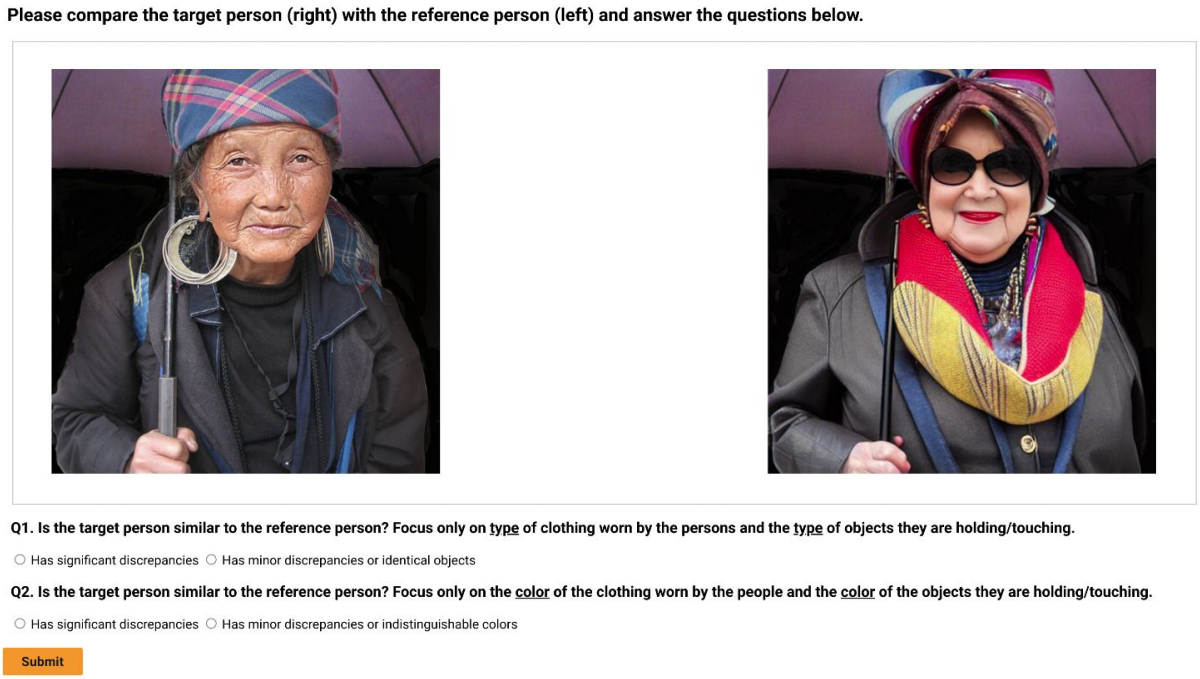


Figure 4.7: Evaluation of *perceived* object and color similarity between original and inpainted images on AMT.

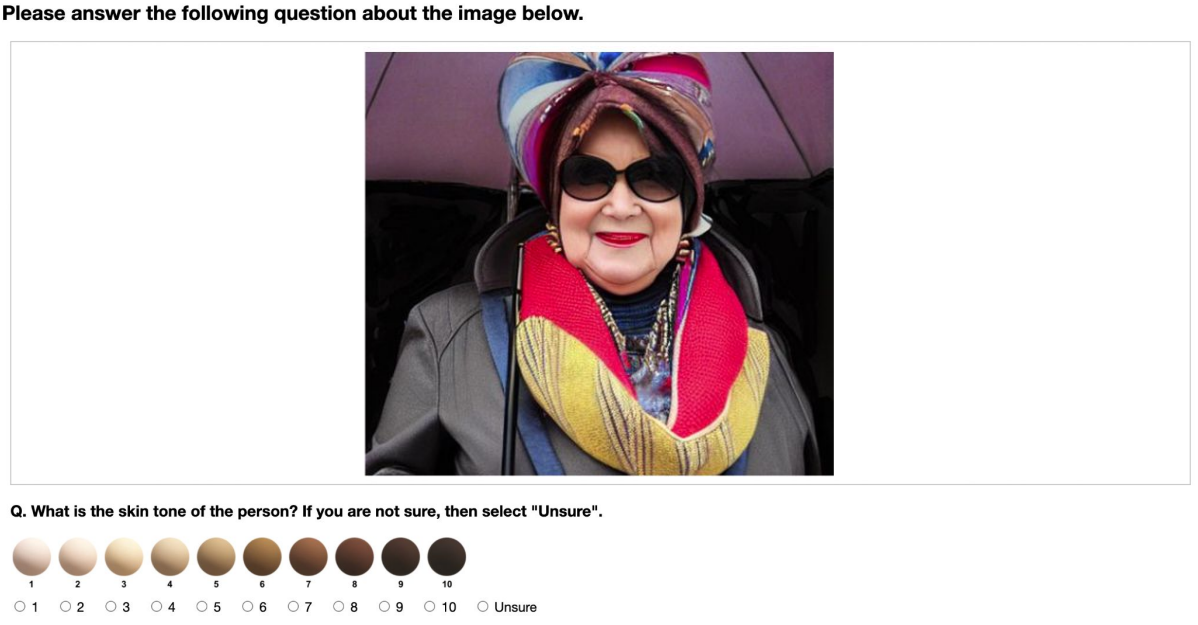


Figure 4.8: Evaluation of *perceived* skin tone using the Monk Skin Tone Scale on AMT.

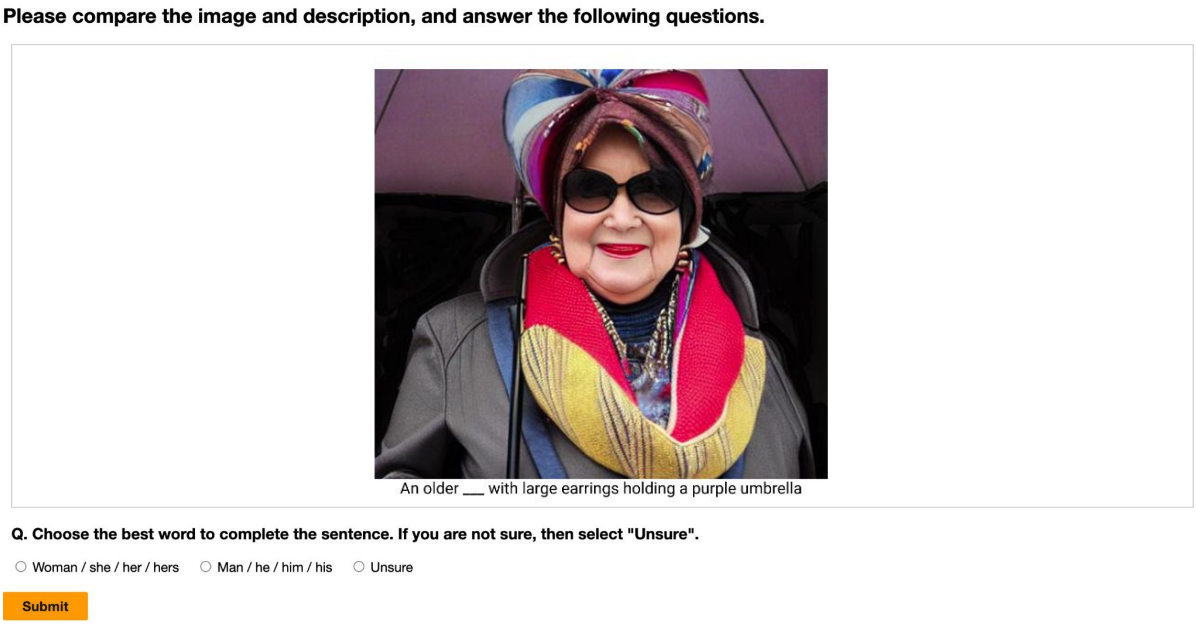


Figure 4.9: Evaluation of *perceived* gender depiction accuracy in inpainted images on AMT.

Chapter 5

Discussion: Relationships to Social Science

The issue of societal bias in AI has been examined from various disciplinary perspectives. In computer vision and related technical fields, including our research, the focus tends to be on identifying and mitigating biases in models and datasets. In contrast, social science research often approaches AI biases through the lens of real-world prejudice and discrimination, exploring the societal structures that influence and are influenced by AI systems. This chapter highlights the differences and connections between these approaches, with examples from both fields.

5.1 Bias in Models and Data: The Technical Perspective

In computer vision, societal bias is primarily studied in terms of its manifestation within datasets and the resulting effects on model outputs. For instance, studies in this domain aim to quantify and reduce disparities in model performance across demographic groups, such as gender or race. These efforts emphasize technical solutions, including data balancing, fairness metrics, and model debiasing algorithms.

5.2 Structural Discrimination: The Social Science Perspective

Social science research often adopts a broader view, analyzing the societal and historical contexts that shape biases in AI. Key discussions include:

The impact of historical discrimination. Research such as [161] argues that long-standing societal discrimination against women and people of color has significantly affected the distribution of data. This skewed data, in turn, contributes to biased AI models. For example, underrepresentation of certain groups in training datasets leads to poorer model performance for those groups, perpetuating inequality.

Amplification of social privilege. Studies like [161, 162] highlight how AI systems often reflect and amplify existing social hierarchies. These systems are frequently designed by and for socially privileged groups, inadvertently reinforcing systemic inequalities. By prioritizing the needs of these groups, AI models risk exacerbating societal disparities, such as unequal access to resources or opportunities.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Professor Yuta Nakashima and Associate Professor Noa Garcia of the Graduate School of Information Science and Technology, Osaka University. Their unwavering support, guidance, and encouragement throughout my doctoral journey have been invaluable, and I am profoundly grateful for their mentorship.

I would also like to extend my heartfelt appreciation to the faculty members of our lab, including Professor Hajime Nagahara, Professor Yuta Nakashima, Associate Professor Noa Garcia, and Associate Professor Hideaki Hayashi. Their insightful feedback and constructive advice have greatly enriched my research and academic experience.

During my doctoral studies, I had the privilege of working as a research scientist intern at Sony AI. I am especially grateful to Dr. Jerone Andrews (Sony AI), Ms. Dora Zhao (Stanford University), Assistant Professor Orestis Papakyriakopoulos (Technical University of Munich), Dr. Apostolos Modas (Sony AI), and Dr. Alice Xiang (Sony AI) for providing me with an intellectually stimulating environment and invaluable collaboration opportunities.

I also had the opportunity to intern at NVIDIA Research, where I was fortunate to work alongside exceptional researchers. I would like to sincerely thank Dr. Ryo Hachiuma, Dr. Chao-Han Huck Yang, Professor Yu-Chiang Frank Wang, Dr. Min-Hung Chen, Dr. Chien-Yi Wang, Dr. Boyi Li, Dr. Yueh-Hua Wu, Dr. Boris Ivanovic, Associate Professor Marco Pavone, and Professor Yejin Choi for their mentorship and collaborative support during my time at NVIDIA Research.

Furthermore, I would like to acknowledge the significant contributions of my co-authors

who have supported me throughout my research endeavors. In particular, I am deeply grateful to Dr. Mayu Otani (CyberAgent, Inc.), Associate Professor Chenhui Chu (Kyoto University), Ms. Yankun Wu (Osaka University), Mr. Tianwei Chen (Osaka University), Mr. Ryan Ramos (Osaka University) for their collaboration and valuable insights.

I am deeply thankful to Professor Takao Onoye and Associate Professor Ittetsu Taniguchi of the Graduate School of Information Science and Technology, Osaka University, who served as my mentor during my undergraduate and master's studies. Their guidance and encouragement laid the foundation for my academic journey and inspired me to pursue a career in research.

Lastly, I would like to thank all those who have contributed to my academic and personal growth during my doctoral studies. To my family and friends, your unwavering support, encouragement, and understanding have been my greatest source of strength throughout this journey. This accomplishment would not have been possible without you. Thank you all.

Reference

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[3] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[5] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[6] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, 2021.

[7] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[9] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*. Springer, 2016.

[10] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 2014.

[11] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *ECCV*. Springer, 2022.

[12] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020.

[13] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Model-agnostic gender debiased image captioning. In *CVPR*, 2023.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[15] Fawaz Sammani and Luke Melas-Kyriazi. Show, edit and tell: a framework for editing image captions. In *CVPR*, 2020.

[16] Robert Wolfe and Aylin Caliskan. Markedness in visual semantic ai. In *FAccT*, 2022.

[17] Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. In *ICCV Workshops*, 2023.

[18] Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. Gender biases in automatic evaluation metrics for image captioning. In *EMNLP*, 2023.

[19] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *FAccT*, 2022.

[20] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. In *ICCV*, 2022.

[21] Jialu Wang, Yang Liu, and Xin Wang. 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*, pp. 1995–2008, 2021.

[22] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, 2018.

[23] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

[24] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

[25] Sakib Mahmud Khan, M Sabbir Salek, Vareva Harris, Gurcan Comert, Eric A Morris, and Mashrur Chowdhury. Autonomous vehicles for all? *Journal on Autonomous Transportation Systems*, 2024.

[26] Zhisheng Chen. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*, 2023.

[27] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[29] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *ICML*, 2021.

[30] Kate Crawford and Trevor Paglen. Excavating AI: The politics of training sets for machine learning. https://excavating.ai, 2019. Accessed: 2021-11-12.

[31] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. In *ECCV*, 2020.

[32] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 547–558, 2020.

[33] Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *FAccT*, 2021.

[34] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *CVPR Workshops*, 2019.

[35] Pierre Stock and Moustapha Cisse. ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In *ECCV*, 2018.

[36] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *ECCV Workshops*, 2018.

[37] William Thong and Cees GM Snoek. Feature and label embedding spaces matter in addressing image classifier bias. In *BMVC*, 2021.

[38] Catherine D'ignazio and Lauren F Klein. *Data feminism*. MIT press, 2020.

[39] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 2019.

[40] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[43] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[44] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *WACV*, 2021.

[45] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *WWW*, 2021.

[46] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, 2004.

[47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[48] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *ICCV*, 2019.

[49] Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. Mitigating gender bias amplification in distribution by posterior regularization. In *ACL*, 2020.

[50] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *ICML*, 2021.

[51] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.

[52] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[53] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004.

[54] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[55] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, Vol. 521, No. 7553, 2015.

[56] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, 1997.

[57] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[58] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NeurIPS*, 2016.

[59] Sunipa Dev and Jeff Phillips. Attenuating bias in word vectors. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

[60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

[61] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. Investigating gender bias in BERT. *Cognitive Computation*, 2021.

[62] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *ACM FAccT*, 2021.

[63] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[64] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, 2020.

[65] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[66] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[67] Xudong Shen, Yongkang Wong, and Mohan Kankanhalli. Fair representation: Guaranteeing approximate multiple group fairness for unknown tasks. *TPAMI*, 2022.

[68] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *AACL*, 2022.

[69] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Gender and racial bias in visual question answering datasets. In *ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[70] Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

[71] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *ICML*, 2020.

[72] Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, and Anupam Datta. Feature-wise bias amplification. In *ICLR*, 2019.

[73] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *NAACL workshop*, 2022.

[74] Dora Zhao, Jerone TA Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. *arXiv preprint arXiv:2210.11924*, 2022.

[75] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, 2022.

[76] Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*, 2019.

[77] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. *arXiv preprint arXiv:2002.08911*, 2020.

[78] Noa Garcia, Yusuke Hirota, Wu Yankun, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *CVPR*, 2023.

[79] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. Imperfect imaGANation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses. *arXiv preprint arXiv:2001.09528*, 2020.

[80] Yi Zhang, Junyang Wang, and Jitao Sang. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *ACMMM*, 2022.

[81] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *ECCV*, 2022.

[82] Ruichen Yao, Ziteng Cui, Xiaoxiao Li, and Lin Gu. Improving fairness in image classification via sketching. In *NeurIPS Workshop*, 2022.

[83] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 2017.

[84] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[85] Fawaz Sammani and Mahmoud Elsayed. Look and modify: Modification networks for image captioning. In *BMVC*, 2019.

[86] Zhen Wang, Long Chen, Wenbo Ma, Guangxing Han, Yulei Niu, Jian Shao, and Jun Xiao. Explicit image caption editing. In *ECCV*, 2022.

[87] Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *ICCV*, 2017.

[88] Alexander C Berg, Tamara L Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. Understanding and predicting importance in images. In *CVPR*. IEEE, 2012.

[89] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 2020.

[90] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

[91] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICLR*. PMLR, 2021.

[92] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[93] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 2020.

[94] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[95] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[96] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019.

[97] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

[98] Dora Zhao, Jerone TA Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. In *ICML*, 2023.

[99] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[100] Sharat Agarwal, Sumanyu Muku, Saket Anand, and Chetan Arora. Does data repair lead to fair models? curating contextually fair data to reduce model bias. In *WACV*, 2022.

[101] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. In *ICCV*, 2023.

[102] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[103] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[104] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *ICLR*, 2023.

[105] Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia (FATE/MM)*, 2020.

[106] Brandon Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. *arXiv preprint arXiv:2305.15407*, 2023.

[107] Jannik Brinkmann, Paul Swoboda, and Christian Bartelt. A multidimensional analysis of social biases in vision transformers. In *ICCV*, 2023.

[108] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical considerations for responsible data curation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[109] Rui-Jie Yew and Alice Xiang. Regulating facial processing technologies: Tensions between legal and technical considerations in the application of illinois bipa. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, p. 1017–1027, 2022.

[110] Benjamin Sobel. A taxonomy of training data: Disentangling the mismatched rights, remedies, and rationales for restricting machine learning. *Artificial Intelligence and Intellectual Property (Reto Hilty, Jyh-An Lee, Kung-Chung Liu, eds.), Oxford University Press, Forthcoming*, 2020.

[111] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[112] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8466–8475, 2018.

[113] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision (ECCV)*, pp. 19–35. Springer, 2016.

[114] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *FAccT*, 2023.

[115] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023.

[116] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? In *EMNLP*, 2022.

[117] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. In *NeurIPS*, 2023.

[118] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

[119] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[120] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2017.

[121] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *CVPR Workshop on Fairness, Accountability Transparency, and Ethics in Computer Vision*, 2019.

[122] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS D&B)*, 2024.

[123] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6957–6966, 2023.

[124] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *CVPR*, 2021.

[125] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 2019.

[126] Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. Contrastive examples for addressing the tyranny of the majority. *arXiv preprint arXiv:2004.06524*, 2020.

[127] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*, 2019.

[128] Zhao Wang and Aron Culotta. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *AAAI*, 2021.

[129] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[130] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

[131] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2019.

[132] Abhishek Mandal, Susan Leavy, and Suzanne Little. Multimodal composite association score: Measuring gender bias in generative multimodal models. *arXiv preprint arXiv:2304.13855*, 2023.

[133] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. *arXiv preprint arXiv:2302.03675*, 2023.

[134] Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. In *ACL*, 2023.

[135] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint arXiv:2209.08891*, 2022.

[136] Eddie L Ungless, Björn Ross, and Anne Lauscher. Stereotypes and smut: The (mis) representation of non-cisgender identities by text-to-image models. In *ACL*, 2023.

[137] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *AIES*, 2023.

[138] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*, 2023.

[139] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

[140] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, 2006.

[141] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022.

[142] Yusuke Hirota, Ryo Hachiuma, Chao-Han Huck Yang, and Yuta Nakashima. From descriptive richness to bias: Unveiling the dark side of generative image caption enrichment. In *EMNLP*, 2024.

[143] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *ICML*, 2019.

[144] Maan Qraitem, Kate Saenko, and Bryan A Plummer. From fake to real (ffr): A two-stage training pipeline for mitigating spurious correlations with synthetic data. *arXiv preprint arXiv:2308.04553*, 2023.

[145] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, 2023.

[146] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. 2023.

[147] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020.

[148] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[149] Amazon Mechanical Turk. Amazon mechanical turk. *Retrieved August*, 2012.

[150] Candice Schumann, Gbolahan O Olanubi, Auriel Wright, Ellis Monk Jr, Courtney Heldreth, and Susanna Ricco. Consensus and subjectivity of skin tone annotation for ml fairness. *arXiv preprint arXiv:2305.09073*, 2023.

[151] Ellis Monk. The monk skin tone scale. 2023.

[152] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *CVPR*, 2016.

[153] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. A step toward more inclusive people annotations for fairness. In *AIES*, 2021.

[154] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022.

[155] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[156] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[157] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[158] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[159] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *EMNLP: system demonstrations*, 2020.

[160] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[161] Catherine D'ignazio and Lauren F Klein. *Data feminism*. MIT press, 2023.

[162] Kate Crawford. The atlas of ai: Power, politics, and the planetary costs of artificial intelligence, 2021.

List of Publications

Journal Publications (related to this thesis)

1. **Yusuke Hirota**, Yuta Nakashima, Noa Garcia. "Societal Bias in Image Captioning: Identifying and Measuring Bias Amplification", *IEICE Transactions on Information and Systems, 2024* (accepted)

2. **Yusuke Hirota**, Yuta Nakashima, Noa Garcia. "Mitigating Gender Bias Amplification in Image Captioning", *IEICE Transactions on Information and Systems, 2024 (conditional acceptance)*

3. **Yusuke Hirota**, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima. "A Picture May Be Worth a Hundred Words for Visual Question Answering", *MDPI Electronics 13, no. 21: 4290. https://doi.org/10.3390/electronics13214290*

4. Yuta Nakashima, **Yusuke Hirota**, Yankun Wu, and Noa Garcia, "Societal bias in vision-and-language datasets and models," *Journal of the Imaging Society of Japan, vol. 62, no. 6, pp. 599–609, Dec. 2023 (doi: https://doi.org/10.11370/isj.62.599)*

International Conference (related to this thesis)

1. **Yusuke Hirota**, Jerone T. A. Andrews, Dora Zhao, Orestis Papakyriakopoulos, Apostolos Modas, Yuta Nakashima, Alice Xiang. "Resampled Datasets Are Not Enough: Mitigating

Societal Bias Beyond Single Attributes", *Conference on Empirical Methods in Natural Language Processing, 2024*

2. **Yusuke Hirota**, Yuta Nakashima, Noa Garcia. "Model-Agnostic Gender Debiased Image Captioning", *IEEE / CVF Computer Vision and Pattern Recognition Conference, 2023*

3. Noa Garcia, **Yusuke Hirota**, Yankun Wu, Yuta Nakashima. "Uncurated Image-Text Datasets: Shedding Light on Demographic Bias", *IEEE / CVF Computer Vision and Pattern Recognition Conference, 2023*

4. **Yusuke Hirota**, Yuta Nakashima, Noa Garcia. "Quantifying Societal Bias Amplification in Image Captioning", *IEEE / CVF Computer Vision and Pattern Recognition Conference, 2022*

International Conference (not related to this thesis)

1. **Yusuke Hirota**, Boyi Li, Ryo Hachiuma, Yueh-Hua Wu, Boris Ivanovic, Yuta Nakashima, Marco Pavone, Yejin Choi, Yu-Chiang Frank Wang, Chao-Han Huck Yang. "The Devil is In the Image Caption Details: Ultrafine Evaluation of Large Vision-Language Models", *Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2025 (submitted)*

2. **Yusuke Hirota**, Min-Hung Chen, Chien-Yi Wang, Yuta Nakashima, Yu-Chiang Frank Wang, Ryo Hachiuma. "SANER: Annotation-free Societal Attribute Neutralizer for Debiasing CLIP", *International Conference on Learning Representations, 2025 (submitted)*

3. **Yusuke Hirota**, Ryo Hachiuma, Chao-Han Huck Yang, Yuta Nakashima. "From Descriptive Richness to Bias: Unveiling the Dark Side of Generative Image Caption Enrichment", *Conference on Empirical Methods in Natural Language Processing, 2024*

4. Tianwei Chen, **Yusuke Hirota**, Mayu Otani, Noa Garcia, Yuta Nakashima. "Would Deep

Generative Models Amplify Bias in Future Models?", *IEEE / CVF Computer Vision and Pattern Recognition Conference, 2024*

5. **Yusuke Hirota**, Yuta Nakashima, Noa Garcia. "Gender and Racial Bias in Visual Question Answering Datasets", *ACM Conference on Fairness, Accountability, and Transparency, 2022*

6. **Yusuke Hirota**, Noa Garcia, Mayu Otani, Chenhui Chu, Yuta Nakashima, Ittetsu Taniguchi, Takao Onoye. "Visual Question Answering with Textual Representations for Images", *IEEE International Conference on Computer Vision Workshop, 2021*

7. **Yusuke Hirota**, Ittetsu Taniguchi, Takao Onoye. "Parallelization of local path planning for high reliable autonomous drones", *International SoC Design Conference, 2020*

Domestic Conference (not related to this thesis)

1. Yusuke Hirota, Noa Garcia, Mayu Otani, Ghenhui Chu, Yuta Nakashima, Ittetsu Taniguchi, Takao Onoye. "How Far Can We Go with Scene Descriptions for Visual Question Answering?", *IPSJ SIG-CVIM: Computer Vision And Image Media*