

Model-Agnostic Gender Debaised Image Captioning

Yusuke Hirota

Yuta Nakashima

Noa Garcia

{y-hirota@is., n-yuta@, noagarcia@}ids.osaka-u.ac.jp

Osaka University



baseline
a young **boy** riding a skateboard

+LIBRA
a young **girl** riding a skateboard



baseline
a **man** riding a wave on a surfboard

+LIBRA
a **woman** catching a wave on a surfboard



baseline
a man wearing a **suit** holding a banana

+LIBRA
a man in a **jacket** holding a banana



baseline
a young boy holding a **baseball bat**

+LIBRA
a young boy holding a plastic **frisbee**

(a) context \rightarrow gender bias mitigation

(b) gender \rightarrow context bias mitigation

Figure 1. Generated captions by a baseline captioning model (UpDn [2]) and LIBRA. We show the baseline suffers from context \rightarrow gender/gender \rightarrow 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.

Abstract

Image captioning models are known to perpetuate and amplify harmful societal bias in the training set. In this work, we aim to mitigate such gender bias in image captioning models. While prior work has addressed this problem by forcing models to focus on people to reduce gender misclassification, it conversely generates gender-stereotypical words at the expense of predicting the correct gender. From this observation, we hypothesize that there are two types of gender bias affecting image captioning models: 1) bias that exploits context to predict gender, and 2) bias in the probability of generating certain (often stereotypical) words because of gender. To mitigate both types of gender biases, we propose a framework, called LIBRA, that learns from synthetically biased samples to decrease both types of biases, correcting gender misclassification and changing gender-stereotypical words to more neutral ones.

1. Introduction

In computer vision, societal bias, for which a model makes adverse judgments about specific population subgroups usually underrepresented in datasets, is increasingly

concerning [4, 6, 7, 11, 17, 22, 41, 43, 52, 57]. A renowned example is the work by Buolamwini and Gebru [7], 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 [49, 58], 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 [10, 15, 24, 42, 64] 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 [5, 8, 18, 44, 51], 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

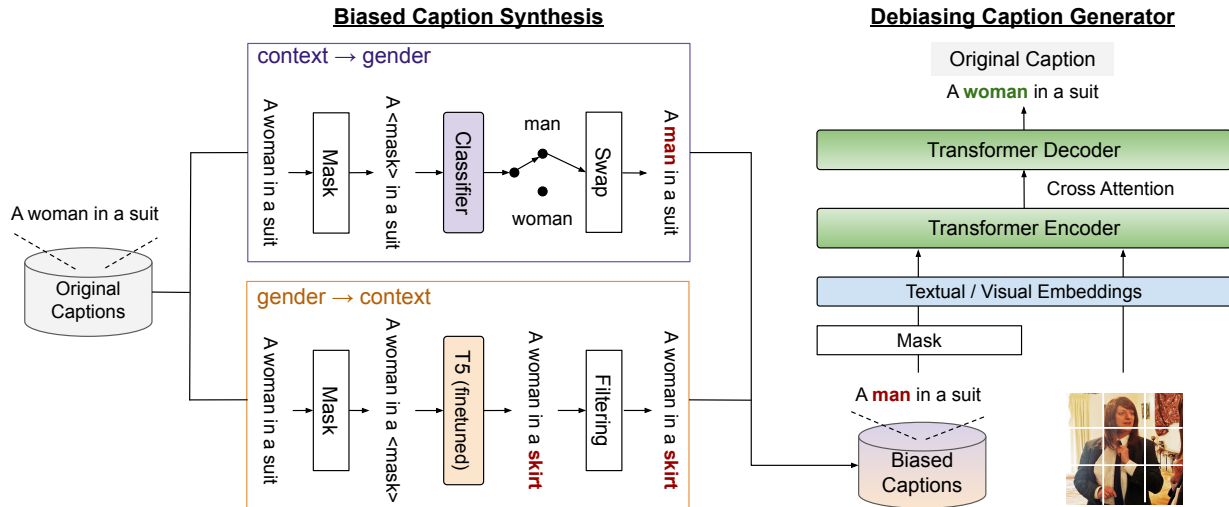


Figure 2. Overview of LIBRA. For the original captions (*i.e.*, ground-truth captions written by annotators), we synthesize biased captions with $\text{context} \rightarrow \text{gender}$ or/and $\text{gender} \rightarrow \text{context}$ bias (Biased Caption Synthesis). Then, given the biased captions and the original images, we train an encoder-decoder captioner, Debiasing Caption Generator, to debias the input biased captions (*i.e.*, predict original captions).

generated words, increasing the probability of predicting certain gender, as shown in Figure 1 (a).

Type 2. $\text{gender} \rightarrow \text{context}$ bias, which increases the probability of generating certain words given the gender of people in an image, as shown in Figure 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 [8], 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 misclassification (reducing $\text{context} \rightarrow \text{gender}$ bias). However, focusing only on decreasing such bias can conversely amplify the other type of bias [18, 51]. For example, as shown in Figure 6, a model trained to correctly predict the gender of a person can produce other words that are biased toward that gender (amplifying $\text{gender} \rightarrow \text{context}$ 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), and 2) Debiasing Caption Generator (DCG), which mitigates bias from synthesized captions (Section 4). Given captions written by annotators, BCS synthesizes biased captions with $\text{gender} \rightarrow \text{context}$ or/and $\text{context} \rightarrow \text{gender}$ biases. DCG is then trained to recover the original caption given a

$\langle \text{synthetic biased caption, image} \rangle$ 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 [8, 18, 44, 66]. This means that DCG can correct gender misclassification caused by the context of the image/words that is biased toward a certain gender, mitigating $\text{context} \rightarrow \text{gender}$ bias (Figure 1 (a)). Also, it tends to change words skewed toward each gender to less biased ones, mitigating $\text{gender} \rightarrow \text{context}$ bias (Figure 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 [30] and SPICE [1]) likely values captions that imitate how annotators tend to describe the gender (*e.g.*, *women posing vs. men standing*).

2. Related work

Societal bias in image captioning In image captioning [2, 62], societal bias can come from both the visual and linguistic modalities [8, 51, 65]. In the visual modality, the image datasets used to train captioning models are skewed regarding human attributes such as gender [14, 37, 59, 65, 66], in which the number of images with men is twice as much as those of women in MSCOCO [26]. Additionally, captions written by annotators can also be biased toward a certain gender because of gender-stereotypical expressions [5, 65], which can be a source of bias from the linguistic modality.

Models trained on such datasets not only reproduce societal bias but amplify it [8, 18, 51, 65]. This phenomenon is demonstrated by Burns *et al.* [8], 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 [19, 20, 44, 46, 50, 53, 55, 56, 60, 63, 66], such as image classification [34] and visual semantic role labeling [61]. For example, Wang *et al.* [53] proposed an adversarial debiasing method to mitigate gender bias amplification in image classification models. In image captioning, Burns *et al.* [8] proposed the Gender equalizer we described in Section 1 to mitigate context \rightarrow gender bias. However, recent work [18, 51] showed that focusing on mitigating gender misclassification can lead to generating gender-stereotypical words and amplifying gender \rightarrow context bias. LIBRA is designed to mitigate bias from the two types of biases.

Image caption editing DCG takes a \langle caption, image \rangle pair as input and debiases the caption. This process is aligned with image caption editing [39, 40, 54] for generating a refined caption. These models aim to correct grammatical errors and unnatural sentences but not to mitigate gender bias. In Section 5.3, we compare DCG with a state-of-the-art image caption editing model [40] and show that a dedicated framework for addressing gender bias is necessary.

3. Biased caption synthesis

Figure 2 shows an overview of LIBRA, consisting of BCS and DCG. This section introduces BCS to synthesize captions with both context \rightarrow gender or/and gender \rightarrow context biases.

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

$$\mathcal{D}_g = F_{GW}(\mathcal{D}), \quad (1)$$

F_{GW} keeps captions that contains either women or men words (*e.g.*, *girl*, *boy*).¹ Therefore, samples in \mathcal{D}_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.

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

²In this paper, we focus on binary gender categories in our framework and evaluation by following previous work [8, 66]. We recognize that the more inclusive gender categories are preferable, and it is the future work.

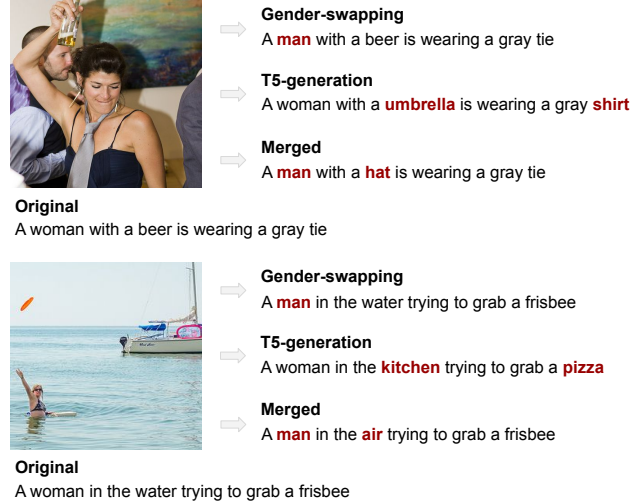


Figure 3. Biased captions synthesized by BCS. Gender-swapping denotes synthesized captions by swapping the gender words (Section 3.1). T5-generation denotes synthesized captions by T5 (Section 3.2). Merged represents biased captions synthesized by applying T5-generation and Gender-swapping (Section 3.3).

3.1. Context \rightarrow gender bias synthesis

Context \rightarrow 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 \rightarrow gender bias. The idea of synthesizing context \rightarrow 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 [3, 45], 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}_g . Let F_{PG} denote the filter that removes captions whose gender is predictable by the sentence classifier. Given $y \in \mathcal{D}_g$, F_{PG} instantiated by first masking gender words and replacing corresponding tokens with the mask token to avoid revealing the gender, following [18]. We denote this gender word masking by $\text{mask}(\cdot)$.

Gender classifier We then train³ gender classifier f_g to predict the gender from masked caption as

$$\hat{g} = f_g(y) = \text{argmax}_g p(G = g | \text{mask}(y)) \quad (2)$$

³Refer to the appendix for training details.

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

$$F_{\text{PG}}(\mathcal{D}_g) = \{y \in \mathcal{D}_g | \hat{g}(y) \neq g\}, \quad (3)$$

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 words (e.g., *man* \rightarrow *woman*) in $y \in F_{\text{PG}}(\mathcal{D}_g)$ results in a biased caption. Letting $\text{swap}(\cdot)$ denote this gender swapping operation, the augmenting set \mathcal{A}_{CG} is given by:

$$\mathcal{A}_{\text{CG}} = \{\text{swap}(y) | y \in F_{\text{PG}}(\mathcal{D}_g)\}. \quad (4)$$

Figure 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*.

3.2. 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 [33] 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 [13].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., $\hat{y} = \text{T5}(y_{\mathcal{M}})$. With this, we can sample an arbitrary number of \hat{y} 's to make a T5-augmented set \mathcal{D}_{T5} as⁵:

$$\mathcal{D}_{\text{T5}} = \{\hat{y} = \text{T5}(y_{\mathcal{M}}) | y \in \mathcal{D}, \mathcal{M} \sim \mathcal{R}\}, \quad (5)$$

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

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

⁵We remove trivial modification that replaces a word with its synonyms based on WordNet [27] and unnatural captions with dedicated classifier. More details can be found in the appendix.

Filtering We then apply a filter to \mathcal{D}_{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). For this, we only use captions in \mathcal{D}_{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 $y \in \mathcal{D}_g$ by a predefined margin δ . Gender filter F_{GF} is given by:

$$F_{\text{GF}}(\mathcal{D}_{\text{T5},g}, \mathcal{D}_g) = \{y' \in \mathcal{D}_{\text{T5},g} | \hat{g}(y') = g, d(y', y) > \delta\}. \quad (6)$$

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

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

$$\mathcal{A}_{\text{GC}} = F_{\text{GF}}(\mathcal{D}_{\text{T5},g}, \mathcal{D}_g). \quad (7)$$

As a result, the synthesized captions contain gender-stereotypical words that often co-occur with that gender as shown in Figure 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. 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:

$$\mathcal{A} = \{\text{swap}(y) | y \in F_{\text{PG}}(\mathcal{D}_{\text{T5},g})\}, \quad (8)$$

which means that the process for synthesizing **context \rightarrow gender** bias in Eqs. (3) and (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 (refer to Merged).

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 [23] that takes an image and text as input and outputs a multi-modal representation. The decoder is a Transformer-based language model [32] that generates text given the encoder's output. The encoder's output is fed into the decoder via a cross-attention mechanism [38].

Training Let $\mathcal{D}^* = \mathcal{A}_{\text{CG}} \cup \mathcal{A}_{\text{GC}} \cup \mathcal{A} = \{(I, y^*)\}$ denote the set of synthetic biased captions where y^* is a biased

Table 1. Dataset construction. Swap denotes synthesized captions by Gender-swapping (Section 3.1). T5 denotes synthesized captions by T5-generation (Section 3.2). Ratio represents the ratio of the number of each type of biased data.

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

caption. When training DCG, given a (I, y^*) 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}_{ce} as

$$\mathcal{L}_{ce} = - \sum_{t=1}^N \log p(y_t | y_{1:t-1}, I, y^*) \quad (9)$$

where p is conditioned on the precedently generated tokens, and I and y^* 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 models and does not require training in captioning models to mitigate gender bias.

5. Experiments

Dataset We use MSCOCO captions [9]. 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 [65]. 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 $|F_{PG}(\mathcal{D}_g)| =$

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 1.

Bias metrics We mainly rely on three metrics to evaluate our framework: 1) **LIC** [18], 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** \rightarrow **context** bias amplification, 2) **Error** [8], which measures the gender misclassification ratio of generated captions. We consider Error to evaluate **context** \rightarrow **gender** bias whereas it does not directly measure this bias (discussed in the appendix), and 3) **BiasAmp** [66], a bias amplification measurement based on word-gender co-occurrence, which is possibly the cause of **gender** \rightarrow **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 [30], CIDEr [48], METEOR [12], and SPICE [1]. While those metrics are widely used to evaluate captioning models, they often suffer from disagreements with human judges [16]. Thus, we also use a reference-free metric, CLIPScore [16], 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 [49], SAT [58], FC [36], Att2in [36], and UpDn [2]) and 2) state-of-the-art Transformer-based models (Transformer [47], OSCAR [25], ClipCap [28], and GRIT [29]). Note that most of the publicly available pre-trained models are trained on the training set of the Karpathy split [21] 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 [8]. Gender equalizer utilizes extra segmentation annotations in MSCOCO [26], which are not always available. The method is not applicable to captioning models that use object-based visual features such as Faster R-CNN [35] because the pre-trained detector’s performance drops considerably for human-masked

Table 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 [31] in their frameworks.

Model	Gender bias ↓			Captioning quality ↑				
	LIC	Error	BiasAmp	BLEU-4	CIDEr	METEOR	SPICE	CLIPScore
NIC [49]	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 [58]	-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 [36]	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 [36]	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 [2]	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 [47]	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 [25]	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 [28]	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 [29]	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

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 [8].

For LIBRA, we use $\delta = 0.2$. The vision-and-language encoder of DCG is Vilt [23], and the decoder is GPT-2 [32]. Unless otherwise stated, we use the combination of biased data composed of T5-generation and Merged in Table 1. We set $\eta = 0.2$ and conduct ablation studies of the settings in Section 5.4 and the appendix.

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 2. We summarize the main observations as follows:

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 1 (b). For example, in the second sample from the right, the baseline, UpDn [2], 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

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

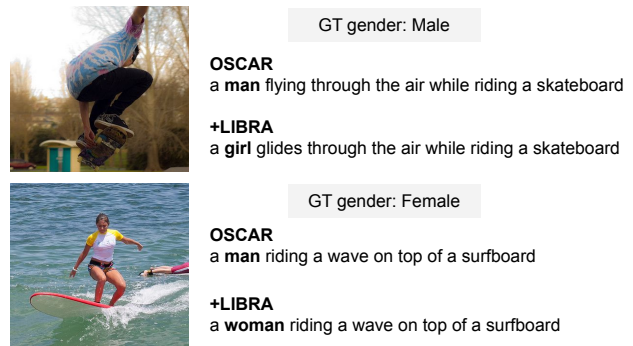


Figure 4. Gender misclassification of LIBRA (Top). Gender misclassification of OSCAR [25] (Bottom). GT gender denotes ground-truth gender annotation in [65].

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


	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	+LIBRA A man posing in front of a white truck
BLEU-4: 79.6 \uparrow METEOR: 43.8 \uparrow	BLEU-4: 47.5 \downarrow METEOR: 33.7 \downarrow
SPICE: 42.9 \uparrow CLIPScore: 74.6 \downarrow	SPICE: 30.8 \downarrow CLIPScore: 76.7 \uparrow

Figure 5. CLIPScore [16] vs. reference-based metrics [1, 12, 30]. 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.

the effectiveness of LIBRA regarding the ability to mitigate **gender \rightarrow context** bias.

LIBRA mitigates context \rightarrow 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 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, skateboard⁸ to increase gender classification accuracy. However, this causes **context \rightarrow gender** bias for images with a gender-context combination rarely seen in the dataset (e.g., women-surfing). In Figure 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 met-

⁷Previous work [5] has shown human annotators possibly annotate gender from context for images without enough cues to judge gender.

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

⁹Surfboard highly co-occur with men in MSCOCO.

Table 3. Comparison with Gender equalizer [8]. Green/red denotes the bias mitigation method mitigates/amplifies bias.

Model	Gender bias \downarrow		Captioning quality \uparrow	
	LIC	Error	SPICE	CLIPScore
NIC+ [49]	1.4	14.6	17.5	69.9
+Equalizer [8]	6.8	7.8	16.8	69.9
+LIBRA	0.4	5.1	18.9	72.7



	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 6. LIBRA vs. Gender equalizer [8].

rics 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 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 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.

5.2. Comparison with other bias mitigation

We compare the performance of LIBRA and Gender equalizer [8] on NIC+ [49], following the code provided by the authors. The results are shown in Table 3. As reported in previous work [18, 51], 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 6, LIBRA predicts the correct gender while not generating gender-stereotypical words. The results of the comparison with Gender equalizer highlight the importance of considering two types of

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

Table 4. Comparison with image caption editing model. Bold numbers represent the best scores in ENT [40] or LIBRA.

Model	Gender bias ↓		Captioning quality ↑	
	LIC	Error	SPICE	CLIPScore
OSCAR [25]	2.4	3.0	24.0	75.8
+ENT [40]	5.7	2.8	21.9	72.8
+LIBRA	0.3	4.6	23.2	75.7

Table 5. Comparison of data used for training DCG. Bold numbers denote the best scores among the types of synthetic datasets.

Model	Synthesis method			Gender bias ↓	
	Swap	T5	Merged	LIC	Error
UpDn [2]	-	-	-	4.7	5.6
+LIBRA	✓	✓	-	2.3	6.2
+LIBRA	-	✓	✓	1.5	4.5
+LIBRA	✓	✓	✓	1.1	5.2
OSCAR [25]	-	-	-	2.4	3.0
+LIBRA	✓	✓	-	-0.8	6.8
+LIBRA	-	✓	✓	0.3	4.6
+LIBRA	✓	✓	✓	0	5.0

biases for gender bias mitigation.

5.3. Comparison with image caption editing model

We compare LIBRA with a state-of-the-art image caption editing model [40] (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 [58] for textual features. The results for OSCAR [25] are shown in Table 4. The complete results are in the appendix. 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.

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 [2] and OSCAR [25]. 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 1. The results are shown in Table 5. The Error score of T5-generation and Merged is consistently the best among the combinations. As for LIC, the results are


Table 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.

Model	Gender bias ↓		Captioning quality ↑	
	LIC	Error	SPICE	CLIPScore
UpDn [2]	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 [25]	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

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}_g and generate words by T5, but without using any filters in Section 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 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.

6. Conclusion

LIBRA  ¹¹ is a model-agnostic framework to mitigate both **context** → **gender** and **gender** → **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 [14], and mechanisms to identify each type of bias.

¹¹This work is partly supported by JST CREST Grant No. JPMJCR20D3, JST FOREST Grant No. JPMJFR2160, JSPS KAKENHI No. JP22K12091, and Grant-in-Aid for Scientific Research (A).

References

- [1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*. Springer, 2016. 2, 5, 7
- [2] 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. 1, 2, 5, 6, 8
- [3] 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. 3
- [4] 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 *AAACL*, 2022. 1
- [5] Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*, 2019. 1, 2, 7
- [6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 1
- [7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM FAccT*, 2018. 1
- [8] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 1, 2, 3, 5, 6, 7
- [9] 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. 5
- [10] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *ICML*, 2020. 1
- [11] Terrance de Vries, Ishan Misra, Changan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *CVPR Workshops*, 2019. 1
- [12] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Workshop on statistical machine translation*, 2014. 5, 7
- [13] 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. 4
- [14] Noa Garcia, Yusuke Hirota, Wu Yankun, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *CVPR*, 2023. 2, 8
- [15] Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022. 1
- [16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP (1)*, 2021. 5, 7
- [17] 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. 1
- [18] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, 2022. 1, 2, 3, 5, 7
- [19] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. Imperfect imagination: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses. *arXiv preprint arXiv:2001.09528*, 2020. 3
- [20] Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. Mitigating gender bias amplification in distribution by posterior regularization. *ACL*, 2020. 3
- [21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 5
- [22] Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *ACM FAccT*, 2021. 1
- [23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICLR*. PMLR, 2021. 4, 6
- [24] Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, and Anupam Datta. Feature-wise bias amplification. In *ICLR*, 2019. 1
- [25] 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. 5, 6, 8
- [26] 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. 2, 5
- [27] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 4
- [28] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 5, 6
- [29] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *ECCV*. Springer, 2022. 5, 6
- [30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002. 2, 5, 7
- [31] 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. 6
- [32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 4, 6

- [33] 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. 4
- [34] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 2017. 3
- [35] 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. 5
- [36] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 5, 6
- [37] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *NAACL-HLT*, 2021. 2
- [38] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 2020. 4
- [39] Fawaz Sammani and Mahmoud Elsayed. Look and modify: Modification networks for image captioning. In *BMVC*, 2019. 3
- [40] Fawaz Sammani and Luke Melas-Kyriazi. Show, edit and tell: a framework for editing image captions. In *CVPR*, 2020. 3, 8
- [41] Xudong Shen, Yongkang Wong, and Mohan Kankanhalli. Fair representation: Guaranteeing approximate multiple group fairness for unknown tasks. *TPAMI*, 2022. 1
- [42] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *NAACL workshop*, 2022. 1
- [43] Pierre Stock and Moustapha Cisse. ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In *ECCV*, 2018. 1
- [44] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *WWW*, 2021. 1, 2, 3
- [45] Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *ICCV*, 2017. 3
- [46] William Thong and Cees GM Snoek. Feature and label embedding spaces matter in addressing image classifier bias. In *BMVC*, 2021. 3
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5, 6
- [48] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDER: Consensus-based image description evaluation. In *CVPR*, 2015. 5
- [49] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 5, 6, 7
- [50] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. In *ECCV*, 2020. 3
- [51] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *ICML*, 2021. 1, 2, 3, 7
- [52] 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. 1
- [53] 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. 3
- [54] Zhen Wang, Long Chen, Wenbo Ma, Guangxing Han, Yulei Niu, Jian Shao, and Jun Xiao. Explicit image caption editing. In *ECCV*, 2022. 3
- [55] 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. 3
- [56] 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. 3
- [57] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019. 1
- [58] 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. 1, 5, 6, 8
- [59] 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 FAccT*, 2020. 2
- [60] Ruichen Yao, Ziteng Cui, Xiaoxiao Li, and Lin Gu. Improving fairness in image classification via sketching. In *NeurIPS Workshop*, 2022. 3
- [61] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*, 2016. 3
- [62] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016. 2
- [63] Yi Zhang, Junyang Wang, and Jitao Sang. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *ACMMM*, 2022. 3
- [64] Dora Zhao, Jerone TA Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. *arXiv preprint arXiv:2210.11924*, 2022. 1
- [65] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, 2021. 2, 3, 5, 6
- [66] 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. 2, 3, 5