Supplementary Material for Model-Agnostic Gender Debiased Image Captioning

Yusuke Hirota Yuta Nakashima Noa Garcia
{y-hirota@is., n-yuta@, noagarcia@}ids.osaka-u.ac.jp
Osaka University

This supplementary material includes:

- Details of BCS (Appendix 1).
- Details of bias metrics (Appendix 2).
- Additional experiments (Appendix 3).
- More visual examples (Appendix 4).
- List of gender-related words (Appendix 5).
- Limitations (Appendix 6).
- Potential negative impact (Appendix 7).

1. Details of BCS

In this section, we provide the details for BCS.

1.1. Training gender classifier

The gender classifier f_{g} is trained on $\mathcal{D}_{\text{g}}$. Specifically, following [6], we split the captions in $\mathcal{D}_{\text{g}}$ into a balanced split with an equal number of samples per female/male, having $66,526$ captions. We use BERT-base [4] 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.

1.2. Finetuning T5

Following the masked language model in [4], 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 $\mathcal{M} = \{m_1, \dots, m_{|\mathcal{M}|}\}$, 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:

$$\mathcal{L}_{mlm} = -\log \prod_{m \in \mathcal{M}} p(y_m | y_{\setminus \mathcal{M}}) \tag{1}$$

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

1.3. Details of T5 masked word generation

To remove trivial modifications, we avoid generating synonyms of the masked tokens by using WordNet [11]. When selecting masked tokens, we chose nouns/verbs/adjectives/adverbs based on POS tagging with NLTK [10].

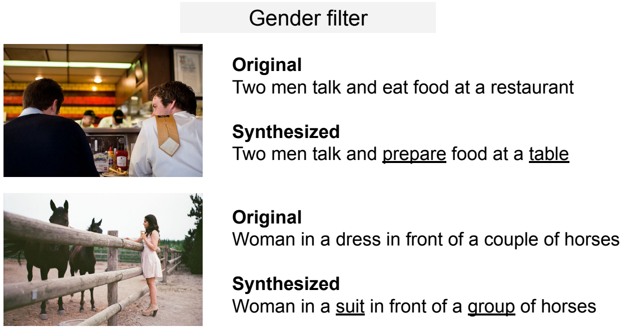

Gender filter
Original
Two men talk and eat food at a restaurant
Synthesized
Two men talk and prepare food at a table
Original
Woman in a dress in front of a couple of horses
Synthesized
Woman in a suit in front of a group of horses

Figure 1. 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 $\mathcal{D}_{\text{T5,g}} \cup \mathcal{D}_g$ to predict whether y is from $\mathcal{D}_{\text{T5,g}}$ or $\mathcal{D}_g$. Let $b \in \{\text{syn}, \text{auth}\}$, prediction $\hat{b}$ is given by:

$$\hat{b} = f_{\text{a}}(y) = \text{argmax}_b \; p(B = b | y) \tag{2}$$

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

$$F_{\text{AF}}(\mathcal{D}_{\text{T5,g}}) = \{y \in \mathcal{D}_{\text{T5,g}} | p(B = \text{auth}|y) > \alpha)\}, \tag{3}$$

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

1.4. Examples of gender/authenticity filter

In Figure 1, 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, in the bottom example, the word *dress* which is skewed toward women is replaced with *suit* which

Table 1. 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

Table 2. Comparison with image caption editing models. Bold numbers represent the best scores in ENT [16] or LIBRA.

	Gender bias ↓		Captioning quality ↑				
Model	LIC	Error	BLEU-4	CIDEr	METEOR	SPICE	CLIPScore
NIC [18]	0.5	23.6	21.9	58.3	21.6	13.4	65.2
+ENT [16]	**-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 [19]	-0.3	9.1	34.5	94.6	27.3	19.2	72.1
+ENT [16]	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 [15]	2.9	10.3	32.2	94.2	26.1	18.3	70.0
+ENT [16]	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 [15]	1.1	5.4	36.7	102.8	28.4	20.2	72.6
+ENT [16]	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 [2]	4.7	5.6	39.4	115.1	29.8	22.0	73.8
+ENT [16]	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 [17]	5.4	6.9	35.0	101.5	28.9	21.1	75.3
+ENT [16]	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 [9]	2.4	3.0	39.4	119.8	32.1	24.0	75.8
+ENT [16]	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 [12]	1.1	5.6	34.8	103.7	29.6	21.5	76.6
+ENT [16]	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 [13]	3.1	3.5	42.9	123.3	31.5	23.4	76.2
+ENT [16]	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

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 1, 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.

2. Details of bias metrics

BiasAmp As for BiasAmp, we also follow the settings presented in [6]. 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 [21], we use words that are strongly related to humans by removing words that do not appear more than 100 times with women/men words.

Table 3. Comparison with DCG without masking input captions. Bold numbers denote the best scores in the DCG with/without masking.

Model	Gender bias ↓		Accuracy ↑	
	LIC	Error	SPICE	CLIPScore
NIC [18]	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 [19]	-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 [15]	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 [15]	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 [2]	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 [17]	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 [9]	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 [12]	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 [13]	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 4. 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
NIC [18]	-	-	-	0.5	23.6
+LIBRA	✓	✓	-	-0.1	7.5
+LIBRA	-	✓	✓	**-0.3**	**5.7**
+LIBRA	✓	✓	✓	-0.2	6.2
SAT [19]	-	-	-	-0.3	9.1
+LIBRA	✓	✓	-	-2.0	6.2
+LIBRA	-	✓	✓	-1.4	**3.9**
+LIBRA	✓	✓	✓	**-2.3**	4.8
FC [15]	-	-	-	2.9	10.3
+LIBRA	✓	✓	-	0.5	6.5
+LIBRA	-	✓	✓	-0.2	**4.3**
+LIBRA	✓	✓	✓	**-0.9**	5.0
Att2in [15]	-	-	-	1.1	5.4
+LIBRA	✓	✓	-	2.0	6.7
+LIBRA	-	✓	✓	-0.3	**4.6**
+LIBRA	✓	✓	✓	**-1.2**	5.5
UpDn [2]	-	-	-	4.7	5.6
+LIBRA	✓	✓	-	2.3	6.2
+LIBRA	-	✓	✓	1.5	**4.5**
+LIBRA	✓	✓	✓	**1.1**	5.2
Transformer [17]	-	-	-	5.4	6.9
+LIBRA	✓	✓	-	**1.5**	6.9
+LIBRA	-	✓	✓	2.3	**5.0**
+LIBRA	✓	✓	✓	2.6	5.8
OSCAR [9]	-	-	-	2.4	3.0
+LIBRA	✓	✓	-	**-0.8**	6.8
+LIBRA	-	✓	✓	0.3	**4.6**
+LIBRA	✓	✓	✓	0	5.0
ClipCap [12]	-	-	-	1.1	5.6
+LIBRA	✓	✓	-	-1.3	6.8
+LIBRA	-	✓	✓	-1.5	**4.5**
+LIBRA	✓	✓	✓	**-1.7**	5.3
GRIT [13]	-	-	-	3.1	3.5
+LIBRA	✓	✓	-	**-0.8**	6.3
+LIBRA	-	✓	✓	0.7	**4.1**
+LIBRA	✓	✓	✓	0	4.8

3. Additional experiments

3.1. Comparison with image caption editing model

We compare LIBRA with a state-of-the-art image caption editing model [16] (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 [19] for textual features.[1] The results are shown in Table 2. 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.

3.2. 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. We can see that masking the input captions of DCG consistently improves the scores on bias

[1] In the original paper, the authors use the captions from AoANet [7]. We use SAT for training ENT as AoANet is trained on Karpathy split [8].

Table 5. 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 [18]	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 [19]	-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 [15]	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 [15]	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 [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
Transformer [17]	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 [9]	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 [12]	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 [13]	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

metrics, which contributes to mitigating two types of biases.

3.3. 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 4. 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 5 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.

4. More visual examples

Bias mitigation by LIBRA Figure 2 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,[2] 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, we show some additional examples of the synthesized captions by BCS. The synthesized captions contain context → gender or/and gender → context biases.

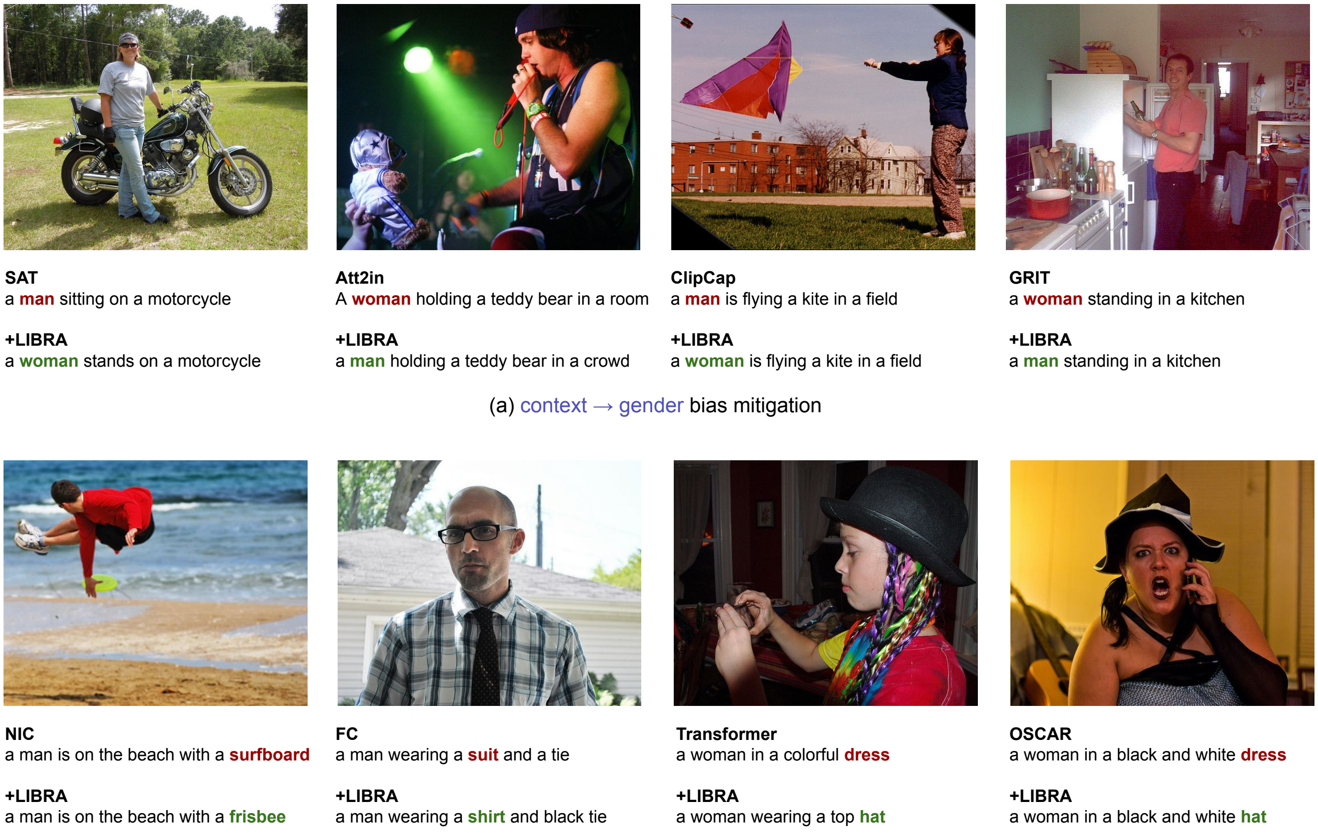
LIBRA vs. human captions The experimental results in the main paper show that LIBRA generates less biased captions than human annotations, resulting in negative LIC scores. Figure 4 shows some visual examples that LIBRA generates more neutral words than human captions. For instance, in the left sample, 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 5, we show the additional examples of the error cases of LIBRA and the state-of-the-art models, OSCAR [9] and GRIT [13]. 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 6, we show the additional examples that LIBRA hurts reference-based metrics by generating words that reduce bias whereas LIBRA does not hurt CLIPScore [5]. For instance, in the left example, the word *little* is skewed toward women in the training set, and LIBRA changed it to *young* which is the less biased word.[3] 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.

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

[3] 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.


SAT
a man sitting on a motorcycle
+LIBRA
a woman stands on a motorcycle
Att2in
A woman holding a teddy bear in a room
+LIBRA
a man holding a teddy bear in a crowd
ClipCap
a man is flying a kite in a field
+LIBRA
a woman is flying a kite in a field
GRIT
a woman standing in a kitchen
+LIBRA
a man standing in a kitchen
NIC
a man is on the beach with a surfboard
+LIBRA
a man is on the beach with a frisbee
FC
a man wearing a suit and a tie
+LIBRA
a man wearing a shirt and black tie
Transformer
a woman in a colorful dress
+LIBRA
a woman wearing a top hat
OSCAR
a woman in a black and white dress
+LIBRA
a woman in a black and white hat

(a) context → gender bias mitigation

(b) gender → context bias mitigation

Figure 2. 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.

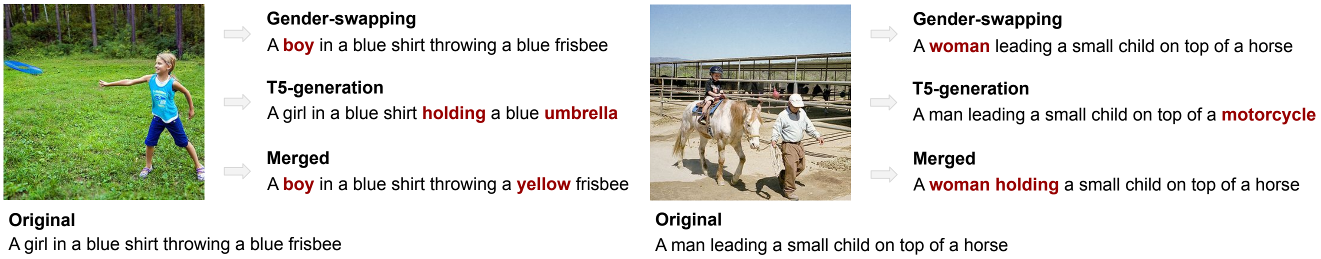

Gender-swapping
A boy in a blue shirt throwing a blue frisbee
T5-generation
A girl in a blue shirt holding a blue umbrella
Merged
A boy in a blue shirt throwing a yellow frisbee
Original
A girl in a blue shirt throwing a blue frisbee
Gender-swapping
A woman leading a small child on top of a horse
T5-generation
A man leading a small child on top of a motorcycle
Merged
A woman holding a small child on top of a horse
Original
A man leading a small child on top of a horse

Figure 3. Biased captions synthesized by BCS.

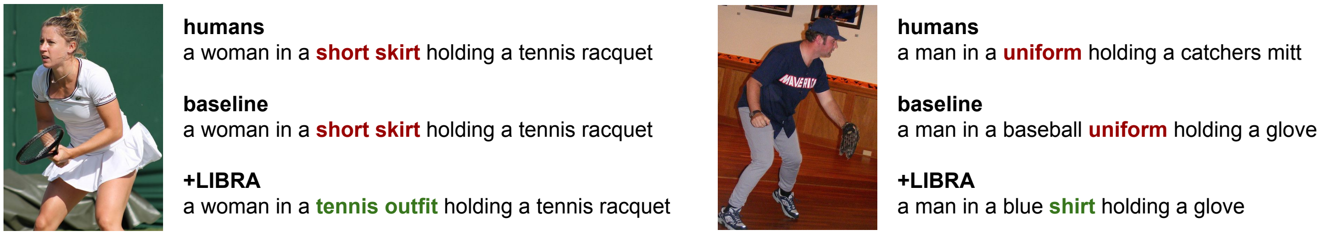
5. List of gender words

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

6. 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



humans
a woman in a **short skirt** holding a tennis racquet

baseline
a woman in a **short skirt** holding a tennis racquet

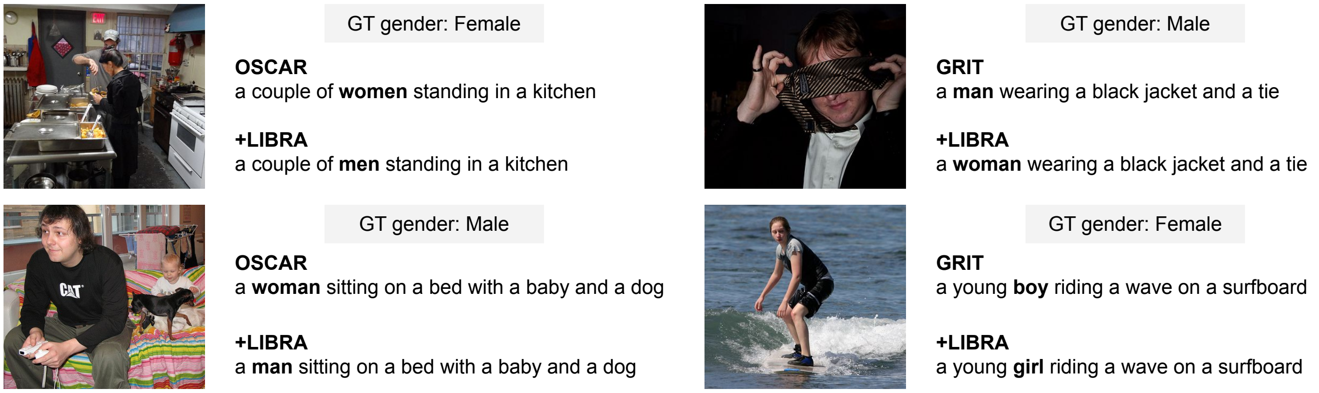
+LIBRA
a woman in a **tennis outfit** holding a tennis racquet

humans
a man in a **uniform** holding a catchers mitt

baseline
a man in a baseball **uniform** holding a glove

+LIBRA
a man in a blue **shirt** holding a glove

Figure 4. Comparison of captions from human annotators, baseline, and LIBRA.



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 5. Gender misclassification of LIBRA (Top). Gender misclassification of OSCAR [9] or GRIT [13] (Bottom). GT gender denotes ground-truth gender annotation in [20].

in the text and tend to be more subtle, for which our bias mitigation method may not work properly.

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.

7. 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.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*. Springer, 2016. 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. 2, 3, 4

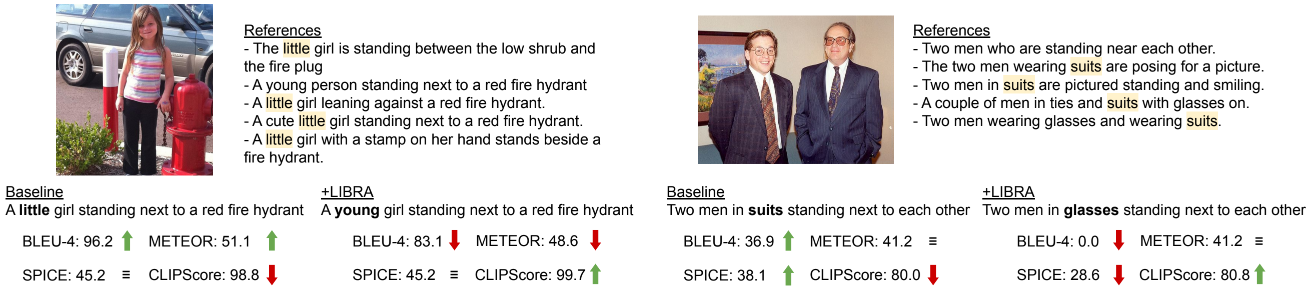
[3] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Workshop on statistical machine translation*, 2014. 7

[4] 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. 1

[5] 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. 4, 7

[6] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, 2022. 1, 2

[7] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019. 3



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 6. CLIPScore [5] vs. reference-based metrics [1, 3, 14]. 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.

[8] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 3

[9] 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. 2, 3, 4, 6

[10] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002. 1

[11] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 1

[12] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 3, 4

[13] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *ECCV*. Springer, 2022. 2, 3, 4, 6

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002. 7

[15] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 2, 3, 4

[16] Fawaz Sammani and Luke Melas-Kyriazi. Show, edit and tell: a framework for editing image captions. In *CVPR*, 2020. 2, 3

[17] 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. 2, 3, 4

[18] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2, 3, 4

[19] 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. 2, 3, 4

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

[21] 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