Supplementary Material for Quantifying Societal Bias Amplification in Image Captioning

This supplementary material includes:

- Experimental details (Appendix A).
- List of gender-related words (Appendix B).
- More visual examples (Appendix C).
- Additional results (Appendix D).
- Potential negative impact (Appendix E).

A. Experimental details

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

A.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 [8].

- **LSTM.** A two-layer bi-directional LSTM [6] 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 [4] 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}.

A.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 [18], 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$_O$ [15] is:

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

$$y_{al} = \mathbb{1}\left[p(a,l) > p(a)p(l)\right] \tag{2}$$

$$\Delta_{al} = \begin{cases} \hat{p}(a|l) - p(a|l) & \text{for DBA}_G \\ \hat{p}(l|a) - p(l|a) & \text{for DBA}_O \end{cases} \tag{3}$$

For DBA$_G$, we use the MSCOCO objects [10] annotated on the images as $\mathcal{L}$ and gender words in the captions as $\mathcal{A}$. For DBA$_O$, we use the MSCOCO objects [10] in the captions as $\mathcal{L}$ and gender annotations [17] as $\mathcal{A}$.

A.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** [16] uses grid-based deep visual features [7] extracted by ResNet [5]. 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.

- **OSCAR** [9] leverages region-based deep visual features [1] extracted by a Faster-RCNN [11]. 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.

B. List of gender-related words

We list the gender-related words that are replaced with the special token when inputting to gender classifiers:

Table 1. 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 LIC$_M = 25$ and LIC $= 0$.

	LSTM			BERT-ft			BERT-pre		
Model	LIC$_M$	LIC$_D$	LIC	LIC$_M$	LIC$_D$	LIC	LIC$_M$	LIC$_D$	LIC
NIC [14]	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 [16]	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 [12]	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 [12]	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 [1]	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 [13]	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 [9]	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+ [2]	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 [2]	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

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$_G$.

C. Visual examples

Here, we show more visual examples that could not be included in the main paper due to space limitations. Figure 1 shows generated captions and their bias score for all the models evaluated in the main paper. Additionally, Figure 2 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 3. The generated caption when the person (man) and the most correlated object (bicycle) are masked still contains a large bias score towards male.

D. Additional results

We compare LIC for race when using different language encoders in Table 1. As with gender bias, the results show that LIC is consistent across different language models.

E. 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.

References

[1] 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

[2] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 2, 4

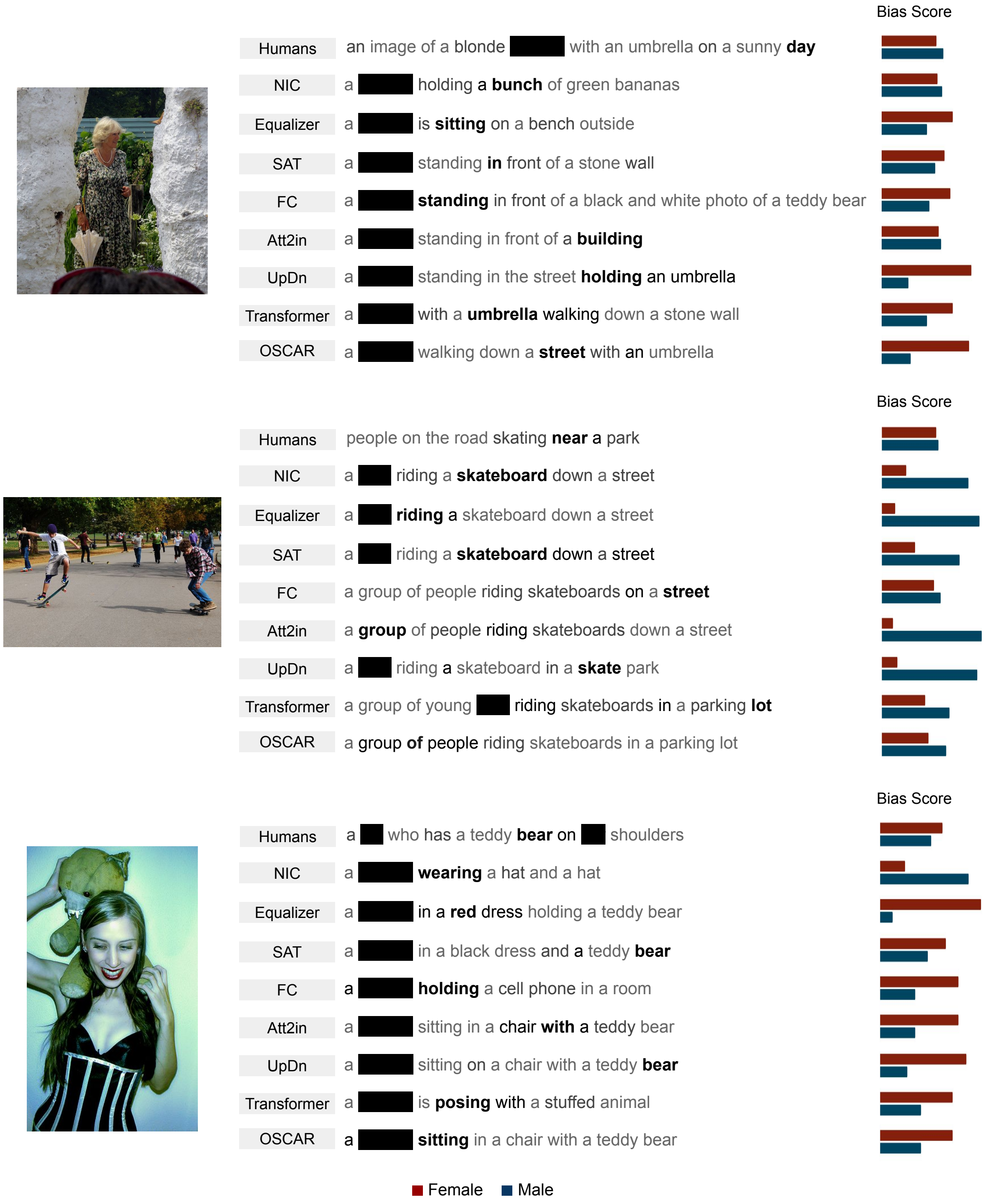
[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 4

[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] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8), 1997. 1

[7] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *CVPR*, pages 10267–10276, 2020. 1



Humans	an image of a blonde [redacted] with an umbrella on a sunny **day**
NIC	a [redacted] holding a **bunch** of green bananas
Equalizer	a [redacted] is **sitting** on a bench outside
SAT	a [redacted] standing **in** front of a stone wall
FC	a [redacted] **standing** in front of a black and white photo of a teddy bear
Att2in	a [redacted] standing in front of a **building**
UpDn	a [redacted] standing in the street **holding** an umbrella
Transformer	a [redacted] with a **umbrella** walking down a stone wall
OSCAR	a [redacted] walking down a **street** with an umbrella

Humans	people on the road skating **near** a park
NIC	a [redacted] riding a **skateboard** down a street
Equalizer	a [redacted] **riding** a skateboard down a street
SAT	a [redacted] riding a **skateboard** down a street
FC	a group of people riding skateboards on a **street**
Att2in	a **group** of people riding skateboards down a street
UpDn	a [redacted] riding a skateboard in a **skate** park
Transformer	a group of young [redacted] riding skateboards in a parking **lot**
OSCAR	a group **of** people riding skateboards in a parking lot

Humans	a [redacted] who has a teddy **bear** on [redacted] shoulders
NIC	a [redacted] **wearing** a hat and a hat
Equalizer	a [redacted] in a **red** dress holding a teddy bear
SAT	a [redacted] in a black dress and a teddy **bear**
FC	a [redacted] **holding** a cell phone in a room
Att2in	a [redacted] sitting in a chair **with** a teddy bear
UpDn	a [redacted] sitting on a chair with a teddy **bear**
Transformer	a [redacted] is **posing** with a stuffed animal
OSCAR	a [redacted] **sitting** in a chair with a teddy bear

Figure 1. 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.

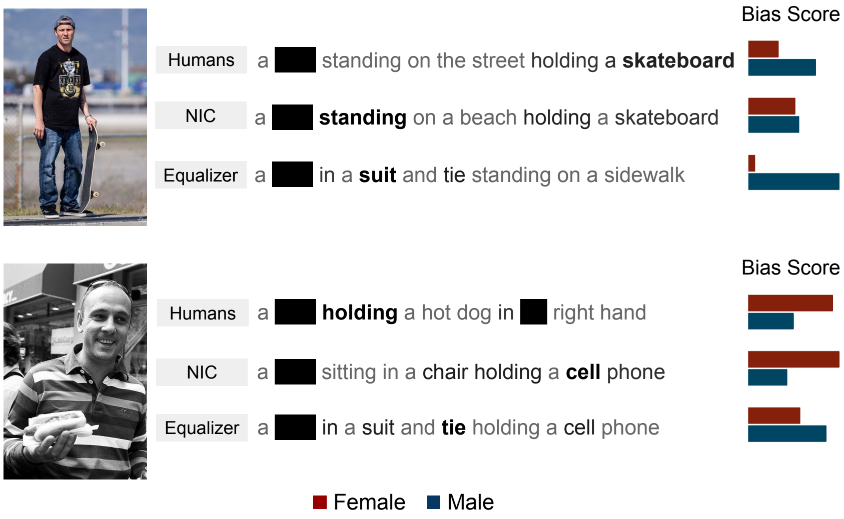


Figure 2. Measuring gender bias in MSCOCO captions [3]. For each caption generated by humans, NIC [14], or NIC+Equalizer [2], we show our proposed bias score for *female* and *male* attributes. The contribution of each word to the bias score is shown in gray-scale (bold for the word with the highest contribution). Gender related words are masked during training and testing.

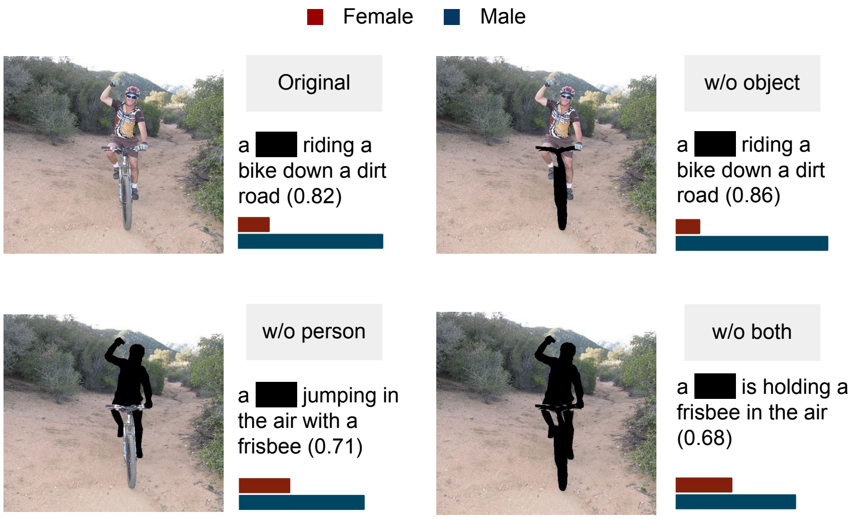


Figure 3. Generated captions and bias scores when images are partly masked.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[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. 1, 2

[10] 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. 1

[11] 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. 1

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

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-

reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

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

[15] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *ICML*, 2021. 1

[16] 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, 2

[17] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, 2021. 1

[18] 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. 1