Mitigating Gender Bias in Captioning Systems

Ruixiang Tang
rxtang@tamu.edu
Department of Computer Science and Engineering, Texas A&M University
College Station, Texas, USA

Mengnan Du
dumengnan@tamu.edu
Department of Computer Science and Engineering, Texas A&M University
College Station, Texas, USA

Yuening Li
liyuening@tamu.edu
Department of Computer Science and Engineering, Texas A&M University
College Station, Texas, USA

Zirui Liu
tradigrada@tamu.edu
Department of Computer Science and Engineering, Texas A&M University
College Station, Texas, USA

Na Zou
nzou1@tamu.edu
Department of Computer Science and Engineering, Texas A&M University
College Station, Texas, USA

Xia Hu
xiahu@tamu.edu
Department of Computer Science and Engineering, Texas A&M University
College Station, Texas, USA

ABSTRACT

Image captioning has made substantial progress with huge supporting image collections sourced from the web. However, recent studies have pointed out that captioning datasets, such as COCO, contain gender bias found in web corpora. As a result, learning models could heavily rely on the learned priors and image context for gender identification, leading to incorrect or even offensive errors. To encourage models to learn correct gender features, we reorganize the COCO dataset and present two new splits COCO-GB V1 and V2 datasets where the train and test sets have different gender-context joint distribution. Models relying on contextual cues will suffer from huge gender prediction errors on the anti-stereotypical test data. Benchmarking experiments reveal that most captioning models learn gender bias, leading to high gender prediction errors, especially for women. To alleviate the unwanted bias, we propose a new Guided Attention Image Captioning model (GAIC) which provides self-guidance on visual attention to encourage the model to capture correct gender visual evidence. Experimental results validate that GAIC can significantly reduce gender prediction errors with a competitive caption quality. Our codes and the designed benchmark datasets are available at https://github.com/datamllab/Mitigating_Gender_Bias_In_Captioning_System.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Social and professional topics** → **Gender**.

KEYWORDS

Fairness, Image Captioning, Gender Bias

ACM Reference Format:
Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating Gender Bias in Captioning Systems. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3442381.3449950

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449950

1 INTRODUCTION

Automatically understanding and describing visual contents is an important and challenging interdisciplinary research topic [12, 18, 19, 26, 40, 43]. Over the past years, thanks to the rapid development of deep learning and huge training data sourced from the web, captioning models have made substantial progress and even surpass humans with regard to several accuracy-based metrics [39]. Although a lot of efforts have been dedicated to improving the overall caption quality, few works consider the potential bias extensively existed among web sourced datasets.

In this work, we investigate the gender bias problem and show that the widely used COCO dataset encodes gender stereotypes in web corpora. Firstly, the occurrence of men in image is significantly higher than women. More precisely, COCO [22] has an unbalanced 1:3 women to men ratio. Second, this gender disparity is even more obvious when considering the gender-context joint distribution. For example, most images about sport co-occur more frequently with men such as 90% of surfboard images only contain male players. As a result, learning models may heavily rely on contextual cues to provide gender identification, e.g., always predicting a person as "woman" when the image is taken in the kitchen, which can lead to incorrect and even offensive gender predictions in domains where unbiased captions are required. The presence of strong priors in the dataset not only results in biased models but also makes it hard to detect the bias learned by models. Due to the i.i.d. nature of randomizing the split between train and test set [1, 14], biased models relying on incorrect visual evidence can still achieve competitive performance on the test set that has similar priors in the training set. This is problematic for validating progress in image captioning since it becomes unclear whether the improvements derive from learning correct visual features or empirical associations.

We present two new COCO splits: COCO-GB V1 and V2 datasets, so as to reveal gender bias learned by captioning models. These two splits are created by reorganizing data distribution such that for each gender, the distribution of image context is very different between train and test set. Specifically, COCO-GB V1 is designed to measure gender bias in existing models systematically. COCO-GB V2 is created to further assess the capabilities of models in gender bias mitigation. Our hypothesis is that models relying on image context to provide gender identification will suffer from a huge gender prediction error on the anti-stereotypical test dataset. In

this way, we can reveal the mismatch between human-intended and model-captured features and quantify gender bias based on the gender prediction outcomes on our new settings.

Equipped with our new benchmark datasets, we evaluate several widely used captioning models. The key observation is that most models learn or even exaggerate gender bias in the training data, which causes high gender prediction errors, especially for women. Another important finding is that commonly used evaluation metrics, such as BLEU [28] and CIDEr [39], mainly focus on the overall caption quality and are not sensitive to gender error. Our experimental results show that a baseline model achieves competitive caption quality scores on these metrics even when it has misclassified 27% of images with women into men. The benchmarking experiments indicate that without extra regularization, captioning models are prone to replicate bias in the dataset. The "high" performance achieved by existing models should be revisited.

In addition to reveal gender bias in learning models, we also seek to mitigate gender bias by guiding the model to capture correct gender features. From the data perspective, a straightforward solution is to train the model on a dataset with equal training samples for each gender. Unfortunately, experimental results indicate that simply balancing image numbers has limited improvement in bias mitigation. From the training perspective, an alternative approach is to increase the loss weight of gender words, which also doesn't achieve a satisfactory result.

To overcome the unwanted bias, we propose a new Guided Attention Image Captioning model (GAIC). GAIC has two complementary streams to encourage the model to explore correct gender features by self-guided supervision. The new training paradigm encourages the model to provide correct gender identification with high confidence when gender evidence in image is obvious. When gender evidence is vague or occluded, GAIC tends to describe the person with gender neutral words, such as "person" and "people." In addition to self-supervised learning, we also consider the semi-supervised scenarios where a small amount of extra supervision is accessible (e.g., person segmentation masks). GAIC training pipeline can seamlessly add extra supervision to accelerate the self-exploration process and further improve gender prediction accuracy. The proposed training paradigm of GAIC is model-agnostic and can be easily applied to various captioning models. Experimental results validate that GAIC can significantly reduce gender prediction errors on our new benchmark datasets, with a competitive caption quality. Results of visualizing attention further prove that GAIC is more inclined to adopt the person's appearance for gender identification. We conclude our contributions as follows:

- Through reorganizing data distribution, we present two new splits COCO-GB V1 and COCO-GB V2 datasets to quantify gender bias learned by captioning models.
- We benchmark several captioning baselines on COCO-GB V1. Experimental results reveal that most models have gender bias problems, leading to high gender prediction errors.
- We propose a new Guided Attention Image Captioning model (GAIC). By self-guided exploration, the model learns to focus on correct gender evidence for gender identification.
- Experimental results demonstrate that the proposed GAIC model can significantly reduce gender prediction errors while at the same time preserving a competitive caption quality.

The rest of this paper is organized as follows. **Section 2** discusses the procedures of dataset creation. **Section 3** presents the benchmarking experiment results on COCO-GB V1. **Section 4** introduces the proposed GAIC and GAIC$_{es}$ model. **Section 5** includes experimental results to verify effectiveness of the proposed method.

2 COCO-GB V1 AND V2

In this section, we present two new splits COCO-GB V1 and V2 to reveal gender bias in learning models.

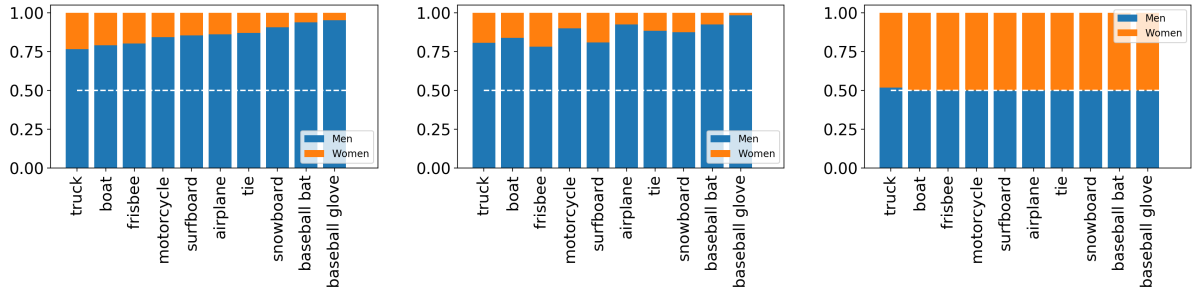
Gender Labeling: Gender identification is not an explicit task for image captioning, hence COCO dataset doesn't particularly label the person's gender. Our first step is to annotate the gender of people in the images. Because many images do not have a clear human face, existing face recognition systems cannot be directly used to annotate the gender. Alternatively, we take inspiration from [16] and make use of the five human-annotated captions available for each image to label the person's gender. Images are labeled as "women" if at least one of the five descriptions contain the female gender words and do not include male gender words. Similarly, images are labeled as "men" if at least one of the five descriptions contain the male gender words and do not include female gender words. Images mentioned both "man" and "woman" are discarded. To improve the label accuracy, we only consider images containing one person and remove images with multiple people.

We then conduct human evaluations on our gender annotations with majority voting from 3 human evaluators. Each evaluator validates a total of 400 test samples from the labeled data, with 200 randomly sampled from each gender. Image is manually labeled from "women," "men," "women & men" (if women and men are included in a single picture) and "discard" (no human appears in the image or gender is indistinguishable). Results show that our gender annotation achieves a high precision, 92% for women, and 95% for men. By analyzing the falsely annotated cases, we find that the gender evidence for these images usually is too vague to distinguish, which makes gender identification far more challenging. (list of gender words, gender annotation examples are shown in Sec. A.1).

COCO-GB V1: COCO-GB V1 is created for systematically evaluating gender bias in existing models. For ease of evaluation, we construct the dataset based on a widely used split proposed by Karpathy et al [19] (Karpathy split), and collect a gender-balanced secret test dataset from its test split. In this way, captioning models trained on Karpathy split can directly evaluate their gender bias on this secret test set without retraining. Previous work [16] also introduces a small dataset with 1:1 women to men ratio to evaluate gender error. However, it does not consider the bias caused by gender-context co-occurrence. In comparison, our proposed secret test dataset reduces the bias in the gender-context distribution and thus can reflect the real performance. To achieve it, we calculate the gender bias rate towards men for 80 COCO object categories by metrics proposed by work [46]:

$$\frac{count(object, men)}{count(object, men) + count(object, women)}, \tag{1}$$

where *men* and *women* refer to images labeled as "men" and "women", $count(object, gender)$ refers to the co-occurrence counts of the object (e.g., Motorcycle and Frisbee) with certain gender. For an object,



(a) (b) (c)

Figure 1: We select several objects in COCO dataset and show their gender distribution in (a) training data, (b) testing data and (c) COCO-GB V1 secret test data. Our key observation is that there is a significant bias in training set, e.g., more than 90% objects have higher probability to co-occur with men. We can find that similar bias also exists in the original test set, while the secret test dataset from COCO-GB V1 has a balanced gender-context distribution.

Model	BLEU-4	CIEDr	METEOR	Gender Error (Original/Secret)	Women			Men		
					correct	wrong	neutral	correct	wrong	neutral
FC	31.4	95.8	24.9	9.7 / 12.6	60.5	14.8	24.7	64.3	10.3	25.4
LRCN	30.0	90.8	23.9	11.3 / 14.0	61.7	16.8	21.6	64.7	11.2	24.0
Att	31.0	95.1	24.8	6.7 / 15.6	53.2	25.1	22.7	62.7	4.4	32.9
AdapAtt	31.1	98.2	25.2	5.5 / 15.3	47.9	26.8	27.4	**75.7**	**3.8**	20.5
Att2in	32.8	102.0	23.3	8.3 / 11.4	61.5	16.3	22.1	70.2	6.5	23.3
TopDown†	34.6	107.6	26.7	7.8 / 8.2	65.5	**9.0**	25.5	63.5	7.4	29.0
NBT†	34.1	105.1	26.2	5.1 / 6.8	**72.3**	9.3	18.3	77.6	4.3	18.0
FC*	33.4	103.9	25.0	9.1 / 13.8	61.6	20.7	17.7	71.9	6.8	21.3
LRCN*	29.4	93.0	23.5	12.1 / 13.4	68.1	11.3	20.6	60.6	15.5	23.9
Att2in*	33.6	106.7	25.7	9.3 / 12.4	61.8	19.7	18.5	73.8	6.2	20.0
TopDown†*	**34.9**	**117.2**	**27.0**	9.1 / 11.3	69.0	15.1	15.8	73.3	7.5	19.1

Table 1: Gender bias analysis on COCO-GB V1 split. We utilize BLEU-4(B-4), CIEDr(C) and METEOR(M) to evaluate captions qualities, all results are generated with beam size 5. Caption quality is obtained from test dataset, and gender bias is evaluated on COCO-GB V1 secret test dataset. † denotes the models that utilize extra grounded information from the Faster R-CNN network. (*) denotes the models that are trained with self-critical loss.

a bias ratio of 0.5 represents that women and men have an equal probability of co-occurring with the object. High/low bias ratio represents that the object is more likely to co-occur with men/women. We obtain the bias ratio of 80 COCO objects in train and test data using Eq. 1 (examples are shown in Fig. 1). The results show that objects in training data have an average bias ratio of 0.65, and 90% of objects are more likely to co-occur with men. A similar bias is also found in the test dataset. We note that the demonstration of bias in COCO is a refection of social bias captured in web.

Evaluating models on the biased test dataset is problematic since models can make use of the contextual cues existed in training data to provide "correct" gender identification. To construct an unbiased test dataset, we utilize a greedy algorithm to select a secret test dataset from the original test split so that each object category has a nearly equal probability of appearing with women or men. The created secret test dataset has 500 images for each gender, and the average bias ratio drops from 0.65 to 0.52. We then validate several existing captioning models on the COCO-GB V1 secret dataset.

COCO-GB V2: This dataset is designed to further assess the model robustness when exposed to novel gender-context pairs at test time. To create the new split, we first sort 80 object categories in COCO dataset according to their gender bias ratios. Unlike creating a balanced test dataset in COCO-GB V1, we start from the most biased object and greedily add selected data into the test set. As a result, the difference has been dramatically enlarged between the gender-context joint distribution of train and test data. Our sampling algorithm guarantees that there are sufficient images from each category during training, but at test time model will face novel compositions of gender-context pairs, e.g., women with the skateboard. The final split has 118,062/5,000/10,000 images in train/val/test respectively (complete gender-context distribution of COCO, COCO-GB V1 and COCO-GB V2 are shown in Sec. A.2).

3 BENCHMARKING EXISTING MODELS

In this section, we benchmark several baselines on the COCO-GB V1 dataset. Models are trained on Karpathy split. We evaluate the caption quality on the original test data and report gender prediction results on both original and COCO-GB V1 test dataset.

Baselines: We compare gender bias across a wide range of models. From the model architecture perspective, we consider models both with and without attention module where "attention" refers to models learn to pay attention to different image regions for caption word generation [31]. For non-attention models, we consider **FC** [30] which initializes LSTM with features extracted directly by

CNN, and **LRCN** [10] which leverages visual information at each time step. For attention models, we select **Att** [43], which firstly applies visual attention mechanism in caption generation. **AdapAtt** [24], which automatically determines when and where to put visual attention. **Att2in** [30], which modifies the architecture of Att, and inputs the attention features only to the cell node of the LSTM. Besides, we also consider models that utilize extra visual grounding information. **TopDown** [2] proposes a novel "top-down attention" mechanism based on Faster R-CNN model [29]. **NBT** [25] generates the sentence "template" with slot locations, and fill slots by an object detection model. All captioning models are end-to-end trainable and use LSTM [17] to output caption text.

Learning Objective and Implementation : We use the cross-entropy loss as the objective function. Besides, FC, LRCN, Att2in, and TopDown are also trained using self-critical loss [30] that uses reinforcement learning to optimize the non-differentiable CIDEr metric. For a fair comparison, all baselines utilize visual features extracted from the ResNet-101 network, NBT and TopDown model adopt extra grounded information from the Faster R-CNN model.

Evaluation Metrics and Results Analysis: In Tab. 1, we report the caption quality as well as gender prediction performance. To evaluate caption quality, we adopt several commonly used metrics, such as BLEU, CIEDr and METOR [9], measure the similarity between machine-generated captions and human-provided annotations. For gender prediction, results are grouped into three circumstances: correct, wrong, and neutral (no gender-specific words are generated). Because of the sensitive nature of prediction for protected attributes (gender words in this work), we emphasize the importance of low error rate [16]. Also, we encourage the model to use gender neutral words in cases where the model has low confidence. Based on the results, we reach the following conclusions.

- In Tab. 1, Gender Error measures the error rates when describing women and men, we find a notable performance gap between the original test dataset (average error rate of 7.7%) and COCO-GB v1 secret test dataset (average error rate of 12.1%), which proves that the original test split indeed underestimates the gender bias learned by models. Tab. 1 also reports the outcome distribution for each gender on the secret test. We observe that the error rate of women is substantially higher than men, the average error rate of all models for women and men is 16.7% and 7.7%, respectively. An interesting finding is that models with the attention mechanism have much higher errors for women (average of 22.6%) compared to non-attention models (average of 15.8%). One possible explanation is that the attention mechanism strengthens models' ability to capture visual concepts; on the other hand, also makes models prone to learn contextual bias. Another finding is that models using extra visual grounding information, such as NBT and TopDown, have a much lower error rate, especially for women (average of 9.1%), which indicates that the extra visual features provided by Faster R-CNN model are unbiased and useful for gender prediction. On the flip side, these models' gender prediction accuracy will highly depend on the features provided by the object detection models.
- Models with a high gender error rate can still get competitive caption quality scores. For example, AdapAtt model obtains a decent caption quality performance on three evaluation metrics but at the same time, has the highest women error rate across all models, which misclassifies 26.8% of images labeled as women into men. The experimental results demonstrate that existing evaluation metrics mainly focus on the overall caption quality and are not sensitive to the gender word error. New evaluation metrics that can effectively lead to high caption quality and low gender prediction errors are still lacking.
- Self-critical loss improves the models' overall caption quality but also significantly amplifies the gender error rate. CIEDr metric obtains an average improvement of 6.2% after training with self-critical loss. However, the error rate of women in FC, Att2in, and TopDown model increases by an average of 5.1% and the error rate of men in LRCN model increases by 4.2%. This result suggests once again that the improvement of overall caption quality cannot guarantee a high accuracy rate of gender identification.

4 THE PROPOSED GAIC FRAMEWORK

Attention mechanism has been widely used in image captioning task, which significantly improves the caption quality [2, 24, 43]. However, our benchmarking experiment shows that when no explicit enforcement is made to infer the gender by only "looking" at the person, models with the attention mechanism are more prone to predict the gender using the context in the image. To overcome the unwanted bias, we propose the Guided Attention Image Captioning model (GAIC) which provides self-guidance on visual attention to encourage the model to utilize correct gender features. More specifically, GAIC considers a situation with no grounded supervision in which the model explores the correct gender evidence by model itself. In this way, the training paradigm of GAIC is model-agnostic and can be easily applied to various captioning models. In addition to self-exploration, we also consider the semi-supervised scenario that uses a small amount of extra supervision to further control attention learning process.

4.1 Caption Generation with Visual Attention

We start by briefly describing the captioning model with attention mechanism. Given an image I and the corresponding caption $y = \{y_1, ..., y_T\}$, the objective of an encoder-decoder image captioning model is to maximize the following loss function:

$$\arg\max_{\theta} \sum_{(I,y)} \log p(y|I;\theta) = \sum_{t=1}^{T} \log p(y_t|y_1, ..., y_{(t-1)}, I), \quad (2)$$

where θ is the trainable parameters of the captioning model. We utilize chain rule to decompose the joint probability distribution into ordered conditionals. Then a recurrent neural network such as LSTM predicts each conditional probability as follows:

$$\log p(y_t|y_1, ..., y_{t-1}, I) = f(h_t, c_t), \quad (3)$$

where f is a nonlinear function that predicts the probability of y_t. h_t is the hidden state of LSTM at t steps. c_t is the visual context vector extracted from image I for predicting t^{th} caption word y_t. We follow the work [43] to compute c_t by:

$$c_t = \sum_{k=1}^{K} \alpha_{t,k} v_k, \quad (4)$$

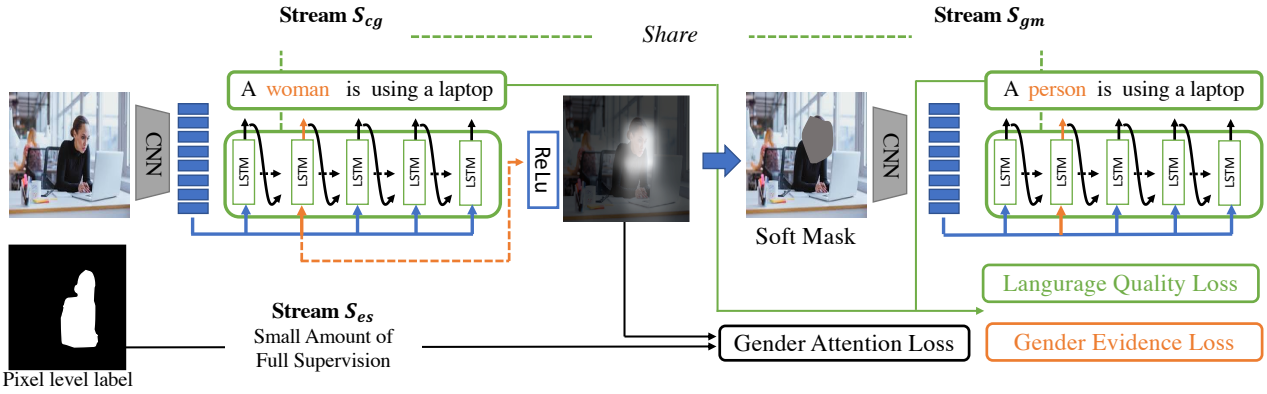


Figure 2: GAIC has two streams of networks that share parameters. Stream S_{cg} finds out regions that help model to classify the gender, and Stream S_{gm} tries to make sure all selected regions are correct gender evidence features. The attention map is online generated and two streams are trained by the Language Quality Loss and Gender Evidence Loss jointly. GAIC$_{es}$ model seamlessly adds a small amount of extra supervision to further refine model attention which denotes as S_{es}.

where $V = \{v_1, ..., v_k\}$,$v_i \in \mathbb{R}^d$ is a set of image features extracted from last convolutional layer of a CNN network. $\alpha = \{\alpha_1, ..., \alpha_T\}$, $\alpha_t \in \mathbb{R}^K$ is the attention weight over features in V. Based on the attention distribution α_t, we extract useful information from image features to obtain context vector c_t. By visualizing the attention weight α_t learned by the model, we are able to analyze on which region of the image does the network focus when generating the t^{th} caption word. On the other hand, we can also modify the generated captions by adding regularization on the α.

4.2 Self-Guidance on Gender Word Attention

To achieve the goal of self-supervision on network attention, we design a two-stream training pipeline. As shown in Fig. 2, GAIC includes caption generation stream S_{cg} and gender evidence mining stream S_{gm}. The two streams share parameters with each other. The purpose of stream S_{cg} is to output high-quality captions and generate attention maps of gender words. The second stream S_{gm} will force the attention generated by S_{cg} to be focused on the correct regions in the image. The two complementary streams guide the attention regions of gender words to be focused on correct features, such as human appearance, and keep away from contextual cues like skateboard and laptop.

For stream S_{cg}, given an image I and ground truth caption $y = \{y_1, ..., y_T\}$, the model outputs a caption supervised by the Language Quality Loss $\mathcal{L}_{lq}$, e.g., the commonly used cross-entropy loss. In addition, we obtain the attention values $\alpha = \{\alpha_t \mid t \in [1, T]\}$ for each caption word where α can be directly obtained from models with attention mechanism like we mentioned in Sec. 4.1. For non-attention models, α can be approximated by the post-hoc interpretation methods, such as Grad-CAM [34] and Saliency Map [35]. In this work, we obtain α directly from an attention model and only consider the attention map of gender words which is denoted as α^g. Such that α^g enables us to visualize the evidence for generating gender words. We then utilze α^g to generate a soft mask that is applied on the original input image. In this way, we obtain I^{*g} using Eq. 5 where I^{*g} represents the image regions apart from the network's attention for gender identification.

$$I^{*g} = I - (I \odot T(\alpha^g)), \tag{5}$$

where $\odot$ denotes the element-wise multiplication. T is a masking function built on a thresholding operation. To make the masking process derivable, we follow the work [21] and use Sigmoid function as an approximation defined in Eq. 6.

$$T(\alpha^g) = \frac{1}{1 + \exp(-w(\alpha^g - E))}, \tag{6}$$

where E is a matrix that all element equal to a threshold value σ. Scale parameter w ensures $T(\alpha_i^g)$ approximately equals to 1 when $T(\alpha_i^g)$ is larger than σ and to 0 otherwise.

We then feed I^{*g} into the second stream S_{gm} and output a new caption. Since I^{*g} already removes gender related features learned by the model, we correspondingly replace the gender words in ground truth caption y into gender neutral words according to the following replacement rules:

$$\begin{aligned} &woman/man(etc) \rightarrow person,\ women/men(etc) \rightarrow people,\\ &boy/girl(etc) \rightarrow child,\ boys/girls(etc) \rightarrow children. \end{aligned} \tag{7}$$

We retrain the model supervised by the new ground truth caption. The loss in S_{gm} is denoted as Gender Evidence Loss $\mathcal{L}_{ge}$, which is defined as follows:

$$\begin{aligned} \mathcal{L}_{ge} = \arg\max_{\theta} \sum_{t=1}^{T} \log p(y_t | y_1, ..., y_{t-1}), I^{*g}),\\ y_t \Leftarrow g_n \ if\ y_t \in g_w \cup g_m, \end{aligned} \tag{8}$$

where y_t denotes the t^{th} caption word, g_n, g_w and g_m represent neutral, female and male gender words, respectively. Loss $\mathcal{L}_{ge}$ can be minimized only when models focus on correct gender features such as human appearance. For a biased model relying on the contextual cues for gender identification, I^{*g} generated by S_{cg} will remove the biased context, e.g., a laptop. As a result, the stream S_{gm} will generate a low-quality caption because of missing the important context features and produce a high loss value for $\mathcal{L}_{ge}$. In such a way, stream S_{gm} forces the stream S_{cg} to capture the correct gender evidence. Besides learning proper gender features, $\mathcal{L}_{ge}$ also encourages the model to predict gender cautiously and use gender neutral words when gender evidence is vague in the

image like I^{*g} (we show an example in Fig. 2). Finally, we combine $\mathcal{L}_{lq}$ and $\mathcal{L}_{ge}$ as the Self-Guidance Gender Discrimination Loss:

$$\mathcal{L}_{self} = \mathcal{L}_{lq} + \mu\mathcal{L}_{ge}, \tag{9}$$

where μ is a weighting parameter to balance two streams. With the joint optimization of $\mathcal{L}_{self}$, the model learns to generate high quality captions as well as focus on correct visual features that contribute to gender recognition.

4.3 Integrating with Extra Supervision

In addition to self-exploration training, we also consider the semi-supervised scenarios where a small amount of extra supervision is added to accelerate the self-exploration process. Based on this idea, we propose the extension of GAIC: GAIC$_{es}$, which can seamlessly integrate extra supervision into the self-exploration training pipeline. More specifically, we utilize the pixel-level person segmentation masks to guide the network to focus on the described person for gender identification. Given the person segmentation masks, GAIC$_{es}$ has another loss, Gender Attention Loss $\mathcal{L}_{ga}$, to further fine-tune the learning process of α^g. Note that the way to generate attention maps of gender words α^g in GAIC$_{es}$ is the same as that in GAIC. $\mathcal{L}_{ga}$ is defined as follows:

$$\mathcal{L}_{ga} = 1 - \sum(\alpha^g \odot (1 - M)), \tag{10}$$

where M denotes the binary pixel-level person segmentation mask which has values 1 in regions of person and 0 elsewhere. Loss $\mathcal{L}_{ga}$ encourages the attention maps of gender words not to exceed the regions of person and thus accelerates the learning process of gender features. The final loss of GAIC$_{\text{es}}$ is defined as follows:

$$\mathcal{L}_{es} = \mathcal{L}_{self} + \eta\mathcal{L}_{ga}, \tag{11}$$

where η is the weighting parameter depending on how much emphasis we want to put on the extra supervision. Since labeling pixel-level segmentation maps are extremely time-consuming and costly, we prefer to use a very small amount of images with external supervision, e.g., 10% in the following experiments. In Fig. 2, we utilize S_{es} to denote the extra supervision stream, and all three streams S_{cg}, S_{gm} and S_{es} share same parameters with each other and can be optimized in an end-to-end manner.

5 EXPERIMENTS

5.1 Experiment Settings and Baselines

In our experiments, GAIC model is built on a widely used captioning model Att [43] that is proved to learn gender bias in the benchmarking experiments. For GAIC$_{\text{es}}$ model, we use 10% images with person segmentation masks as the extra supervision. All models are trained with stochastic gradient descent using adaptive learning rate algorithms. We use early stopping on BLEU score to select the best model and set $\sigma = 10, \mu = 0.1, \eta = 0.05$ in our experiments. We compare GAIC with the following three debiasing baselines:

- **Balanced:** A subset is selected from the training data, which has a balanced gender ratio (4,000 images for each gender). We then fine-tune Att model on this new dataset. This baseline helps investigate the correlation between gender ratio and gender bias.
- **UpWeight:** We also conduct an experiment with up-weighting the loss value of gender words' during training. To this end, we label the gender words position for each ground truth caption and multiply a constant value on the loss of gender words (we set the constant value to 5 and 10, which is denoted as UpWeight-5 and UpWeight-10). UpWeight forces the model to accurately predict gender words. However, unlike the Gender Evidence Loss, UpWeight does not encourage the model to predict gender neutral words when gender evidence is too vague to identify.
- **Pixel-level Supervision (PixelSup):** As a variant of GAIC$_{es}$ model, we remove self-exploration streams and directly use Eq. 10 to fine-tune model attention with 10% extra data, force gender words attention α_g not to exceed the person segmentation masks.

5.2 Evaluation Metrics

We consider the following metrics to evaluate captioning models:

- **Gender Accuracy:** Unlike the traditional binary gender classification, in this work, gender prediction results are grouped into correct, wrong, and neutral three categories. Due to the sensitive nature of prediction for protected attributes, reducing the error rate is most important. On this basis, the captioning model should predict gender cautiously and outputs neutral words when gender is too vague to distinguish.
- **Gender Divergence:** We also expect models to treat each gender fairly. To this end, inspired by [16], we utilize Cosine Distance to measure the outcome similarity between Women and Men. For a fair system, it should have similar outcomes (correct/wrong/neutral rate) across different groups, which resembles to the fairness definition proposed in Equality of Odds [15]. Lower divergence indicates that Women and Men have a similar distribution of outcomes and can be considered as more fair.
- **Attention Correctness:** To measure whether attention focuses on the correct regions, we compare the attention maps of gender words with the person segmentation masks. We adopt two evaluation metrics to calculate the similarity: Pointing Game [44], which measures whether the point with highest attention value falls in the person segmentation masks, and Attention Sum [23], which calculates the sum of attention weights that fall in the person regions. For the two metrics, high outcome value represents the high similarity and thus can be consider as more correct.

5.3 Experimental Results

GAIC: Tab. 2 reports the gender prediction results on COCO-GB V1. We observe that GAIC significantly improves the gender prediction performance compared to the base model Att. The gender accuracy of women increases from 53.2% to 62.0%, and its error rate reduces from 25.1% to 16.9%. The gender accuracy of men increases from 62.7% to 77.3%, and its error rate reduces from 4.4% to 4.2%.

UpWeight: UpWeight-10 model obtains the highest accuracy of both women and men. However, it also causes the highest error rate for two genders, which is unacceptable. UpWeight-5 obtains a comparable gender correct rate compared with GAIC but has a notable higher error rate. The UpWeight model shows that up-weighting the loss value of gender words will proportionally increase both the correct and error rate and dramatically reduce the neutral rate.

Balanced: An interesting finding is that there is no substantial difference between the Balanced model and Att model, a similar

Model	CIDEr	METOR	Woman			Men			Divergence
			correct	wrong	neutral	correct	wrong	neutral	
Att	95.1	24.8	53.2	25.1	22.7	62.7	4.4	32.9	0.063
Balanced	93.9	24.9	54.3	24.3	21.4	69.7	7.4	22.8	0.032
UpWeight-5	94.1	24.5	60.3	22.4	17.3	73.2	8.1	18.7	0.028
UpWeight-10	93.5	24.2	**70.6**	26.4	2.9	**81.4**	10.5	8.9	0.028
PixelSup	92.0	24.5	57.3	21.1	21.6	68.3	8.6	23.1	0.032
GAIC	93.7	24.6	62.0	16.9	21.1	77.3	4.2	18.5	0.024
GAIC$_{es}$	94.6	24.7	64.1	**13.1**	22.8	75.3	**4.0**	20.7	**0.011**

Table 2: Gender bias analysis on COCO-GB V1 split. GAIC and GAIC$_{es}$ significantly improves the gender prediction performance compared to the baselines. Although UpWeight-10 model obtains the highest accuracy of both women and men, it also causes an unacceptable high error rate for two genders. For fairness evaluation, GAIC and GAIC$_{es}$ obtain the lowest divergence, which indicates that women and men have a similar outcome distribution and thus can be considered as more fair.

(a) Attention Sum

Accuracy	Women	Men	Average
Baseline	25.5	21.2	23.4
Balanced	25.0	21.4	23.2
Upweight-10	26.7	23.3	25.0
PixelSup	30.0	28.1	29.0
GAIC	27.4	24.3	25.6
GAIC$_{es}$	**32.5**	**28.5**	**30.1**

(b) Point Game

Accuracy	Women	Men	Average
Baseline	64.8	57.4	61.1
Balanced	66.2	59.6	62.9
Upweight-10	66.2	59.3	62.8
PixelSup	67.2	60.5	63.9
GAIC	67.2	61.2	64.2
GAIC$_{es}$	**67.8**	**61.5**	**64.7**

Table 3: Attention Correctness on COCO-GB V1 split. We adopt Point Game [44] and Attention Sum [23] as evaluation metrics. Pointing Game measures the probability that point with highest attention value falls in the person segmentation masks. Attention Sum calculates the sum of attention weights that fall in the regions of person. The high outcome represents the correct of the model attention. Results show that GAIC and GAIC$_{es}$ model can significantly improve attention correctness.

result is found by [5], which indicates that models learn gender bias mainly from bias in gender-context co-occurrence. Simply balancing the gender ratio of training samples thus has a limited improvement. A more efficient way is to eliminate gender-context bias. However, we emphasize that balancing the distribution for every gender-context pair in large-scale data is extremely difficult and guaranteeing complete gender-context decoupling is infeasible.

PixelSup: We observe that PixelSup model obtains sensible improvements, which indicates that directly adding supervision on model's attention can control caption word prediction.

GAIC$_{es}$: Compared to GAIC, GAIC$_{es}$ obtains consistently better performance. We empirically find that adding extra supervision accelerates the self-exploration process and makes training process more stable. To formalize the notion of fairness, we calculate the outcome divergence. GAIC and GAIC$_{es}$ have the lowest divergence, which indicates that women and men have a similar outcome distribution and thus can be considered as more fair.

Experiments on COCO-GB V2 have similar trends in COCO-GB V1 (detailed analysis of COCO-GB V2 results is in Sec. C). Compared to COCO-GB V1, all baselines in COCO-GB V2 obtain a worse gender prediction result. This is mainly because the unseen gender-context pairs in the test dataset increase the prediction difficulty. In comparison, our proposed GAIC and GAIC$_{es}$ model obtain a comparable performance on COCO-GB V1 dataset, which proves the robustness of the proposed self-exploration training strategy.

Caption Quality: Besides gender accuracy, we also expect the model to generate linguistically fluent captions. We use METEOR(M) and CIDEr(C) to evaluate caption quality. Results are shown in Tab. 2, GAIC and GAIC$_{es}$ only cause a minor performance drop compared to the baseline (drop from 95.1 to 94.6 on METROR and drop from 24.8 to 24.7 on CIDEr). In Fig. 3, examples show that the sentences generated by GAIC and GAIC$_{es}$ are linguistically fluent with more correct gender descriptions.

Attention Correctness: To measure whether the model focuses on the correct gender features, we calculate the similarity between attention maps of gender words and the person segmentation masks. Quantitative results are shown in Tab. 3. We observe that GAIC and GAIC$_{es}$ receive consistent improvement over the Att model and all variants on two evaluation metrics: Pointing Game and Attention Sum, which indicates that the proposed models tend to focus on the described person for the gender word prediction. We also show qualitative comparisons in Fig. 3. We observe that the baseline Att model utilizes biased visual features and thus makes incorrect gender prediction, e.g., predicting a woman as a boy based on a tennis ball. In comparison, GAIC and GAIC$_{es}$ learn to concentrate on the regions of the described person for gender word prediction. In addition to correct gender prediction, proposed models learn to use gender neutral words when gender is too vague to distinguish. We put more visualizing attention results and discussions in Sec. E.

6 RELATED WORK

Gender Bias in Dataset: The issue of gender bias has been studied in a variety of AI domains [3, 5–8, 11, 13], especially in natural language processing domain [16, 20, 33, 36, 38, 45, 46]. It has been reported that many popular language data resources contain gender bias [32, 37, 41, 42]. As a result, the language models trained on the dataset may learn the bias and thus has preference or prejudice toward one gender over the other. Several studies have shown that

Figure 3: Qualitative comparison of baseline and our proposed models.

the language models can learn and even amplify the gender bias in the dataset [5, 16, 36, 46]. For instance, in visual semantic role labeling task, it shows that cooking images in the training set are twice that more likely involve females than males, and the model trained on the dataset further amplifies the disparity to five times during inference [46]. In regard to image captioning task, several initial attempts have been devoted to study gender bias in the captioning dataset. An early work analyzes the stereotype in Flicker30k dataset and mentions the gender-related issues [27]. Another work demonstrates that caption annotators in COCO dataset tend to infer a person's gender from the image context when gender cannot be identified, e.g., annotators tend to describe snowboard players as "man" even gender is usually vague for the snowboard player [16].

Vision and Language Bias: Bias in captioning models can be grouped into vision and language bias [31]. The vision bias refers to the use of wrong visual evidence for caption generation. While the language bias refers to capture unwanted language priors in human annotations [31]. For example, the phrase "in a suit" often follows the word "man" in the training data. RNN's recurrent mechanism enables caption models to learn this language prior and predict the phrase "in a suit" after the word "man" regardless of the image contents. In this work, we notice that gender words are usually mentioned at the beginning of the sentence (on the average at position 2 with an average sentence length of 9), and the words inferred before gender words, such as "a" and "the," do not have any gender preference, which indicates that gender words cannot affect by the predicted words. Hence the gender bias problem we studied in this work should mainly come from the vision part.

Mitigating Gender Bias: Few initial attempts have been made to overcome gender biases in captioning models. One solution is to make image captioning a two-step task [4]. Firstly, an object detection network locates and recognizes the people in the image. Then a language model combines this extra grounded information with image embeddings to generate the final captions. For this method, gender prediction accuracy highly depends on the object detection model. Compared to the two-step training strategy, GAIC has an end-to-end architecture and thus does not rely on other grounded information. In another work [16], the authors alleviate the gender bias by modifying the training images and loss functions. However, their approach requires a high-quality person segmentation mask for each training data, which is costly and infeasible for many large-scale datasets. In comparison, the proposed self-guided attention mechanism can work without extra supervision and thus greatly expands the application scenarios.

7 CONCLUSIONS AND FUTURE WORK

In this paper, two novel COCO splits are created for studying gender bias problem in image captioning task. We provide extensive baseline experiments for benchmarking different models, training strategies, as well as a comprehensive analysis of our new datasets. The experimental results indicate that many captioning models utilize contextual cues for gender identification, leading to undesirable gender prediction errors, especially for women. To overcome this unwanted bias, we propose a new training framework GAIC, which can significantly reduce gender bias by self-guided supervision. Experimental results validate that the proposed models learn to focus on the described person for gender identification.

As an initial attempt to approach the potential gender bias in captioning systems, we emphasize that utilizing the gender prediction

accuracy to quantify gender bias is not enough, evaluation metrics that can effectively lead to high caption quality and low gender bias are still lacking. Understanding and evaluating the caption bias from the causal inference perspective is a promising research field and would be exploded in our future research.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their helpful comments. The work is in part supported by NSF IIS-1900990, and NSF IIS-1939716. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4971–4980.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6077–6086.

[3] Emily M Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[4] Shruti Bhargava. 2019. *Exposing and correcting the gender bias in image captioning datasets and models*. Ph.D. Dissertation.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[6] Shikha Bordia and Samuel Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 7–15.

[7] Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the Origins of Bias in Word Embeddings. In *International Conference on Machine Learning*. 803–811.

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[9] Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*. 85–91.

[10] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[11] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* (2020).

[12] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1473–1482.

[13] Joel Escudé Font and Marta R Costa-jussà. 2019. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 147–154.

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut Learning in Deep Neural Networks. *arXiv preprint arXiv:2004.07780* (2020).

[15] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[16] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*. Springer, 793–811.

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[18] MD Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys (CSUR)* 51, 6 (2019), 118.

[19] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[20] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science* 65, 7 (2019), 2966–2981.

[21] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. 2018. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9215–9223.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[23] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017. Attention correctness in neural image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[24] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 375–383.

[25] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7219–7228.

[26] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632* (2014).

[27] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E Papalexakis, and Amit K Roy-Chowdhury. 2018. Webly supervised joint embedding for cross-modal image-text retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*. 1856–1864.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[30] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7008–7024.

[31] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4035–4045.

[32] Karen Ross and Cynthia Carter. 2011. Women and news: A long and winding road. *Media, Culture & Society* 33, 8 (2011), 1148–1165.

[33] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 8–14.

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[35] K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. (2014).

[36] Pierre Stock and Moustapha Cisse. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 498–512.

[37] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).

[38] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3003–3008.

[39] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[40] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[41] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *International AAAI Conference on Weblogs and Social Media*. USA, 454–463.

[42] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 602–608.

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural

image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[44] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2018. Top-down neural attention by excitation backprop. *International Journal of Computer Vision* 126, 10 (2018), 1084–1102.

[45] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 629–634.

[46] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2979–2989.

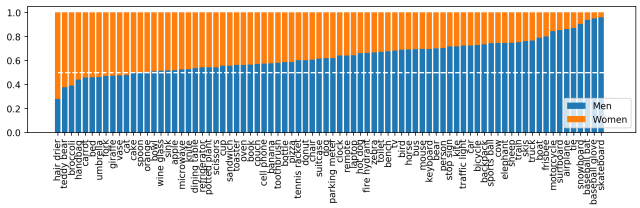


Figure 4: Gender-context distribution of COCO train

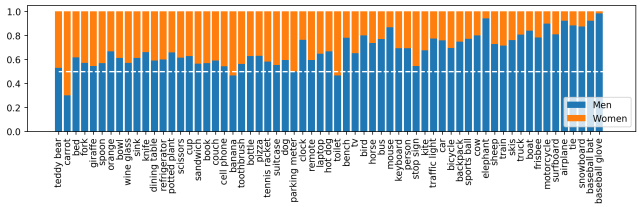


Figure 5: Gender-context distribution of original COCO test

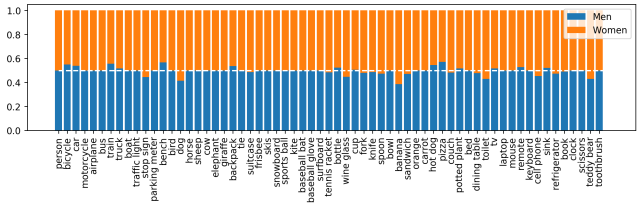


Figure 6: Gender-context distribution of COCO-GB V1 test

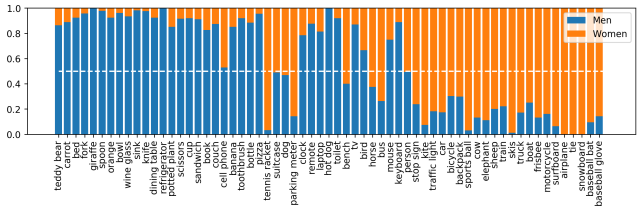


Figure 7: Gender-context distribution of COCO-GB V2 test

A COCO-GB DATASET

A.1 Gender Annotation

We show the gender word list in Tab. 4. Gender words are selected based on the word frequency in COCO dataset. We delete the gender words that appear less than ten times in training data. Word "woman" and "man" are the most frequent gender-specific word and account for more than 60% of the total gender-specific words.

female word (g_w)	woman, women, girl, sister, daughter, wife, girlfriend
male word (g_m)	man, men, boy, brother, son, husband, boyfriend
gender neutral word (g_n)	people, person, human, baby

Table 4: Gender words list. We label the caption based on the gender word appeared in the sentences.

Some gender annotation examples are shown in Fig. 8. We label an image as "women" when at least one sentence mentioned female words and label an image as "men" when at least one sentence mentioned male words. Images that both mention male words and female words are discarded.

We conduct human evaluations on our gender annotations with majority voting from 3 human evaluators. Each evaluator validates a total of 400 test samples, with 200 randomly sampled from each gender. Results show that our gender annotation achieves high precision, 92% for women, and 95% for men. Fig. 8 (c) shows an example of failure. The gender label based on captions is Men. However, evaluators label the image as "discard" since the gender is actually vague in the image. By analyzing the examples of failure, we observe that most conflicts occur when gender evidence is vague or occluded. In these cases, COCO annotators tend to provide gender identification based on context cues or social stereotypes.

A.2 Gender-Contex Joint Distribution

In Fig. 4-7, we show the gender-context joint distribution of COCO training dataset (Karpathy split), COCO testing dataset (Karpathy split), COCO-GV V1 secret testing dataset, COCO-GB V2 testing dataset. We list the objects in order of bias rate in COCO training dataset. For COCO-GB dataset, we choose 63 representative objects from a total of 80 objects and delete objects that do not occur in the corpus with sufficient frequency to be include.

The Fig. 4 shows that most objects in training data are more likely to co-occur with men. In Fig. 5, we observe that a similar bias is also found in the test dataset. Hence directly evaluating the models on the biased test dataset might underestimate the gender bias learned by models. In comparison, we observe that COCO-GB V1 has a balanced gender-context joint distribution. Each object has an almost equal probability of occurring with men and women. COCO-GB V2 has a different distribution that gender-context joint distribution is opposite to the training set, which makes the gender prediction more difficult. Models relying on image context to provide gender identification will suffer from a huge gender prediction error on the anti-stereotypical test dataset.

B MORE ON IMPLEMENTATION DETAILS

Benchmarking Baselines: All the baselines mentioned in Sec. 3 use visual features extracted from the fourth layer of ResNet-101. All baselines except for NBT, TopDown, Att, and AdaptAtt are implemented in a same open-source framework [1]. We directly use the caption results from a web source [2] to measure gender prediction performance. For NBT, TopDown, Att, and AdaptAtt models, we implement models based on the paper and make sure that the model's caption score is close to the results reported in work. For all models, we evaluate caption quality by the official COCO evaluation tool [3].

[1] https://github.com/ruotianluo/self-critical.pytorch
[2] https://github.com/LisaAnne/Hallucination
[3] https://github.com/tylin/coco-caption

The woman in the kitchen is holding a huge pan.
A chef carrying a large pan inside of a kitchen.
A woman is holding a large pan in a kitchen.
A woman cooking in a kitchen with counters.
A woman cooking in her kitchen with a black pan.

(a)

A person sitting at a table in a room.
A man who is sitting at a table.
A man in a news room sits in front of a camera.
A man sitting a desk in front of a TV.
person is sitting at a desk with a television behind him

(b)

An explorer or adventurer hikes in a remote area.
a snowboarder sliding down a mountain with wind.
A man riding a snow board down a snow covered slope.
This is a person with a mask on in a sandy place.
A Man on large open area covered with snow.

(c)

Figure 8: (a): An image is labeled as women. (b): An image is labeled as men. (c) An image is labeled as men. However, the gender evidence is actually occluded and our human evaluators label the image as "discard".

Model	C	M	Woman			Men			D
			correct	wrong	neutral	correct	wrong	neutral	
Baseline	98.2	27.2	51.6	28.3	20.1	77.9	**4.9**	17.1	0.094
Balanced	97.5	27.3	57.9	25.5	26.6	71.1	11.5	17.4	0.034
UpWeight-10	95.8	26.9	**72.2**	26.1	1.7	**86.1**	11.7	2.1	0.023
PixelSup	96.8	27.1	54.2	25.1	20.5	76.4	6.2	17.2	0.062
GAIC	97.8	26.9	67.1	18.0	14.9	68.9	10.7	20.3	0.008
GAIC$_{es}$	98.1	27.0	69.1	**15.2**	15.7	71.4	8.1	20.5	**0.007**

Table 5: Gender bias analysis on COCO-GB V2 split.

Debiasing Baselines: Here we discuss the implementation details of baselines mentioned in Sec. 5.1. For baseline Balanced, we randomly select a subset from the original training set which contains 4,000 images for each gender. Balanced model is obtained by fine-tuning the Att model on this selected dataset with 5 epochs. For UpWeight model, we first train the Att model with normal loss. Then we upweight the loss value of gender words and continue to train the model for another 1 epoch. For model PixelSup, we first train the Att model with normal loss. Then we utilize Eq. 10 to further fine-tune the attention learning process on the 10% extra supervision data with one more epoch.

GAIC Model: For GAIC model, we adopt the two-streams pipeline to train the model. In the experiments, we find $\mu = 0.1$ performs best, large μ value will influence caption qualiy, small μ value cannot efficiently mitigate gender bias. For GAIC$_{es}$ model, we fine-tune the dataset with extra human instance segmentation annotations, and set $\mu = 0.1$ and $\eta = 0.05$. Like μ value, we empirically find that large η value will significantly impact caption quality. Hence we choose a relatively small value $\eta = 0.05$ to train GAIC$_{es}$ model.

C EXPERIMENTAL RESULTS ON COCO-GB V2

Compared to COCO-GB V1, all baselines in COCO-GB V2 obtain a higher error rate, especially for women (average increase of 3.25%). GAIC model improves the gender prediction accuracy of woman from 51.6% to 67.1% and reduces the error rate of women from 28.3% to 18.0%. Although the UpWeight-10 model obtains the highest accuracy of both women and men, it causes an unacceptably high error rate for two genders. There is no substantial difference between the Balanced model and baseline model, a similar trend has been found in COCO-GB V1. Compared to GAIC, GAIC$_{es}$ obtains consistently better performance. For fairness evaluation, we compare different model's gender divergence. GAIC and GAIC$_{es}$ obtain the lowest divergence, which indicates that models treat each gender in a fair manner. For caption quality, GAIC and GAIC$_{es}$ only cause a minor performance drop compared to the baseline model, which proves the robustness of the self-exploration training strategy.

D MORE QUALITATIVE RESULTS

We show more visualizing attention results in Fig. E. We observe that the baseline Att model tends to utilize context features to predict gender and thus makes incorrect gender prediction, e.g., predicting a woman as a man based on a tie. In comparison, GAIC and GAIC$_{es}$ learn to use the features on the regions of the described person for gender prediction. The result also shows that when gender evidence is vague, GAIC and GAIC$_{es}$ model tend to use neutral gender words, such as "person" to describe the person.

E MORE ON ETHICS CONCERN

In the COCO dataset, people in the images are not asked to identify their biological gender. Thus we emphasize that we are not classifying biological sex or gender identity, but rather outward gender appearance. In this work, we follow the settings used in previous work to define three gender categories: male, female, and gender-neutral (when the image is blurry or the gender in the image is ambiguous) based on visual appearance. Gender labels are determined based on the annotators' descriptions.

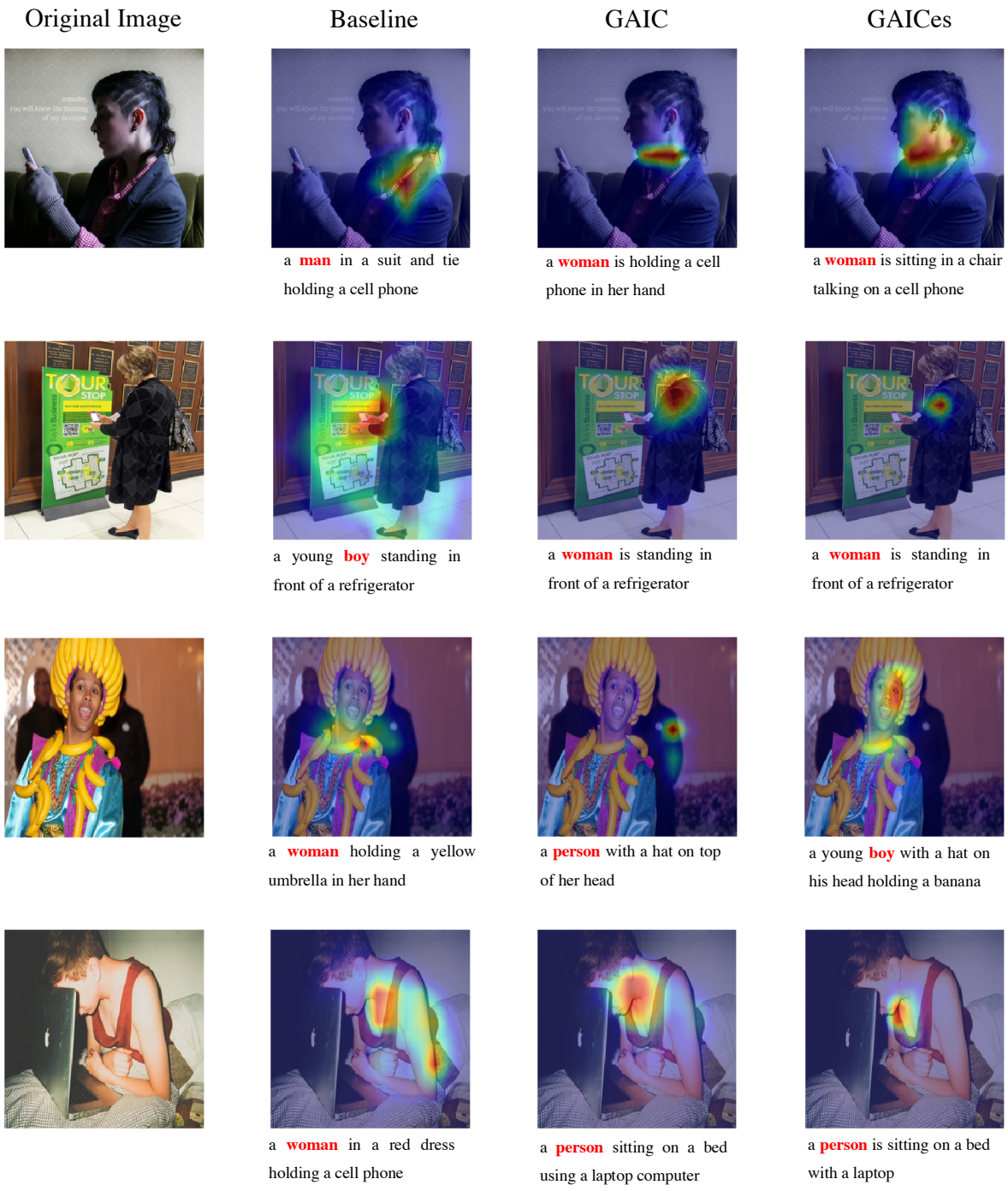


Figure 9: Qualitative comparison of baselines and our proposed model. At the top, we show success cases that our proposed modes correctly predict the gender and utilize proper visual evidence. The bottom case shows that when gender evidence is vague, our model tends to use neutral gender words, such as "person" to describe the gender of the person.