ImageCaptioner^2: Image Captioner for Image Captioning Bias Amplification Assessment

Eslam Abdelrahman[1], Pengzhan Sun[2*], Li Erran Li[3], Mohamed Elhoseiny[1]

[1]King Abdullah University of Science and Technology (KAUST)
[2]National University of Singapore
[3]AWS AI, Amazon

Abstract

Most pre-trained learning systems are known to suffer from bias, which typically emerges from the data, the model, or both. Measuring and quantifying bias and its sources is a challenging task and has been extensively studied in image captioning. Despite the significant effort in this direction, we observed that existing metrics lack consistency in the inclusion of the visual signal. In this paper, we introduce a new bias assessment metric, dubbed ImageCaptioner^2, for image captioning. Instead of measuring the absolute bias in the model or the data, ImageCaptioner^{2}pay more attention to the bias introduced by the model w.r.t the data bias, termed bias amplification. Unlike the existing methods, which only evaluate the image captioning algorithms based on the generated captions only, ImageCaptioner^{2}incorporates the image while measuring the bias. In addition, we design a formulation for measuring the bias of generated captions as prompt-based image captioning instead of using language classifiers. Finally, we apply our ImageCaptioner^{2}metric across 11 different image captioning architectures on three different datasets, i.e., MS-COCO caption dataset, Artemis V1, and Artemis V2, and on three different protected attributes, i.e., gender, race, and emotions. Consequently, we verify the effectiveness of our ImageCaptioner^{2}metric by proposing AnonymousBench, which is a novel human evaluation paradigm for bias metrics. Our metric shows significant superiority over the recent bias metric; LIC, in terms of human alignment, where the correlation scores are 80% and 54% for our metric and LIC, respectively. The code and more details are available at *https://eslambakr.github.io/imagecaptioner2.github.io/*.

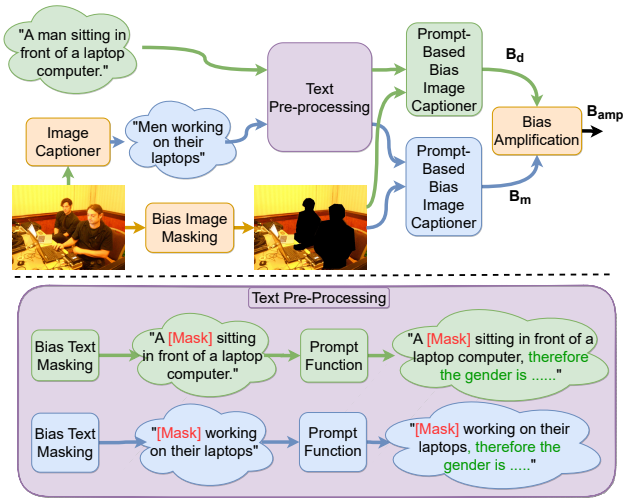

Figure 1: An abstract overview for our metric pipeline. The GT stream is in green, and the prediction stream is in blue. Given an input image, the associated GT-caption, and the predicted caption, a text pre-processing module is utilized to refine the captions before feeding them to the prompt-based image captioners. The text pre-processing module performs two main functionalities: 1) Masking the protected-attribute indicative words. 2) Appending the prompt template to the caption. Finally, bias-amplification module is utilized to measure the model bias B_m w.r.t the GT bias B_d.

Introduction

Most deep learning (DL) benchmarks (Deng et al. 2009) (Lin et al. 2014) (Cireşan et al. 2011) (Kuznetsova et al. 2020) (Liao, Xie, and Geiger 2022) are designed to rank different architectures based on accuracy, neglecting other aspects such as fairness. Recently, measuring the bias and understanding its sources have attracted significant attention due to models' social impact (Alvi, Zisserman, and Nellåker 2018) (De Vries et al. 2019) (Khan and Fu 2021) (Stock and Cisse 2018) (Thong and Snoek 2021) (Wang et al. 2022) (Yang et al. 2020) (Du et al. 2022) (Schick, Udupa, and Schütze 2021). For instance, image captioners may learn shortcuts based on correlation, which inevitably suffers from unreliable associations between protected attributes, e.g., gender, and visual or textual clues (Zhao et al. 2017; Wang et al. 2019).

Recent efforts focus on estimating model bias, driven by the fact that more than balanced data is needed to create unbiased models (Wang et al. 2019). Bias (Zhang et al. 2023; Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017) is characterized by the model's representation of different subgroups when generating the supergroup, such as assessing whether it equally depicts men and women in images of people. The primary cause of bias stems from spurious correlations captured during training. we are interested in the spurious correlation found within ordinary daily scenes,

*This work was done during internship in KAUST.

which typically do not evoke bias concerns. Accordingly, BA (Zhao et al. 2017), DBA (Wang and Russakovsky 2021), and LIC (Hirota, Nakashima, and Garcia 2022) measure the model bias w.r.t the bias in the training dataset. In other words, they emphasize the bias introduced by the model regardless of the bias in the ground-truth dataset. Therefore, detecting and measuring the bias is the first step toward a reliable and accurate image captioner.

BA (Zhao et al. 2017) and DBA (Wang and Russakovsky 2021), calculate the co-occurrence frequency of protected bias attribute and detected objects in the picture or some specific words in the caption as a bias score. An apparent limitation of these evaluations is that only limited visual or language contexts are used for computing the statistics. Consequently, several learnable bias measurements have been proposed recently (Zhao, Wang, and Russakovsky 2021; Hirota, Nakashima, and Garcia 2022; Wang et al. 2019). The basic assumption of this line of work is that if the bias attribute category, e.g., gender, can be inferred through context even if the attribute feature does not exist, then it indicates that the model is biased. For instance, the recent work, LIC (Hirota, Nakashima, and Garcia 2022), estimates the bias in image captioning models using trainable language classifiers, e.g., BERT(Devlin et al. 2018).

Consequently, measuring the bias in image captioning is notably challenging due to the task's multimodal nature. Even though significant progress has been achieved in quantifying the bias and its amplification in image captioning models, it remains to be investigated how to properly measure the bias amplified by the image captioning model over the dataset from both the vision and language modalities.

The existing bias metrics quantify the bias only from the language side, where the image stream is omitted while measuring the bias. Discarding the visiolinguistic nature of the task in the metric may result in misleading conclusions. Addressing these limitations, we propose a novel bias metric termed ImageCaptioner^2, which incorporates the visiolinguistic nature of the image captioning task by utilizing an image captioner instead of a language classifier.

In addition to filling the multi-modalities gap in the existing bias metrics, ImageCaptioner^2 better matches the underlying pre-training model, by proposing a prompt-based image captioner, i.e., the training objective is text generation. In contrast, the existing bias metric, e.g., LIC, utilizes a language classifier trained for a different objective, i.e., classification objective; this leads to a potential under-utilization of the pre-trained model we assess. In other words, our proposed metric provides textual prompts that bring the evaluation task closer to the pre-training task.

Experimentally, we evaluate our metric through comprehensive experiments across 11 different image captioning techniques on three different datasets, i.e., MS-COCO caption dataset(Lin et al. 2014), ArtEmis V1 (Achlioptas et al. 2021), and ArtEmis V2 (Mohamed et al. 2022), and on three different protected attributes, i.e., gender, race, and emotions. In addition, we introduce consistency measures to judge which learnable metric is more consistent against classifiers variations: 1) Introducing conflict-score to count the number of mismatches between different classifiers. 2) Introducing Ranking Consistency, which utilizes Pearson correlation. In comparison to LIC, we show that based on the consistency measures, our method is more consistent than LIC, where the correlation ratio is 97% and 92%, respectively. Consequently, we verify the effectiveness of our ImageCaptioner^2 metric by proposing AnonymousBench, depicted in Figure 2, which is a novel human evaluation paradigm for bias metrics. Our metric shows significant superiority over the recent bias metric; LIC, in terms of human alignment, where the correlation scores are 80% and 54% for our metric and LIC, respectively.

Our contributions can be summarized as follows:

- We develop a novel bias metric called ImageCaptioner^2, depicted in Figure 1. To the best of our knowledge, ImageCaptioner^2 is the first metric that is formulated based on the visiolinguistic nature of the image captioning while quantifying the bias.
- We propose a prompt-based image captioner that better matches the underlying pre-training model we assess.
- Propose AnonymousBench to verify the effectiveness of our ImageCaptioner^2 metric, depicted in Figure 2.
- We propose a new scoring function that matches the objective of measuring the bias.
- We apply our metric across 11 different image captioning techniques on three different datasets, i.e., MS-COCO caption dataset, Artemis V1, and Artemis V2, and on three different attributes, i.e., gender, race, and emotions.

Revisiting Fairness Metrics

In recent years, there have been rapid efforts toward designing robust fairness metrics. Nonetheless, all the existing metrics suffer from severe limitations that make them far from being reliable; therefore, in this paper, we take one step toward designing a more robust and reliable metric. In this section, we first cover the different taxonomies of the existing fairness metrics in image captioning. Then, we dissect them to show their limitations, which raises several questions.

Taxonomy of Fairness Metrics

We classify the fairness metrics into three categories:

Source-Agnostic Vs. Source-Identification metrics. Women also Snowboard (Hendricks et al. 2018), and GAIC (Tang et al. 2021) assume access to the bias attribute in the predicted captions while measuring the misclassification rate. Accordingly, such metrics can not identify the bias source, which is crucial in designing a debiasing technique. In contrast, methods that rely on a pre-trained language model to determine whether the model is biased or not, such as (Zhao, Wang, and Russakovsky 2021), LIC (Hirota, Nakashima, and Garcia 2022), and (Wang et al. 2019), or the methods that rely on calculating the co-occurrences, i.e., BA (Zhao et al. 2017) and DBA (Wang and Russakovsky 2021), have the capability of determining the source of the bias.

Learnable Vs. Non-Learnable metrics. (Zhao, Wang, and Russakovsky 2021) studies the racial bias, i.e., lighter and darker, by utilizing a dedicated classifier to predict the race based on the predicted caption. In addition, LIC (Hirota, Nakashima, and Garcia 2022), and (Wang et al. 2019) train additional language classifiers to measure the bias in the data

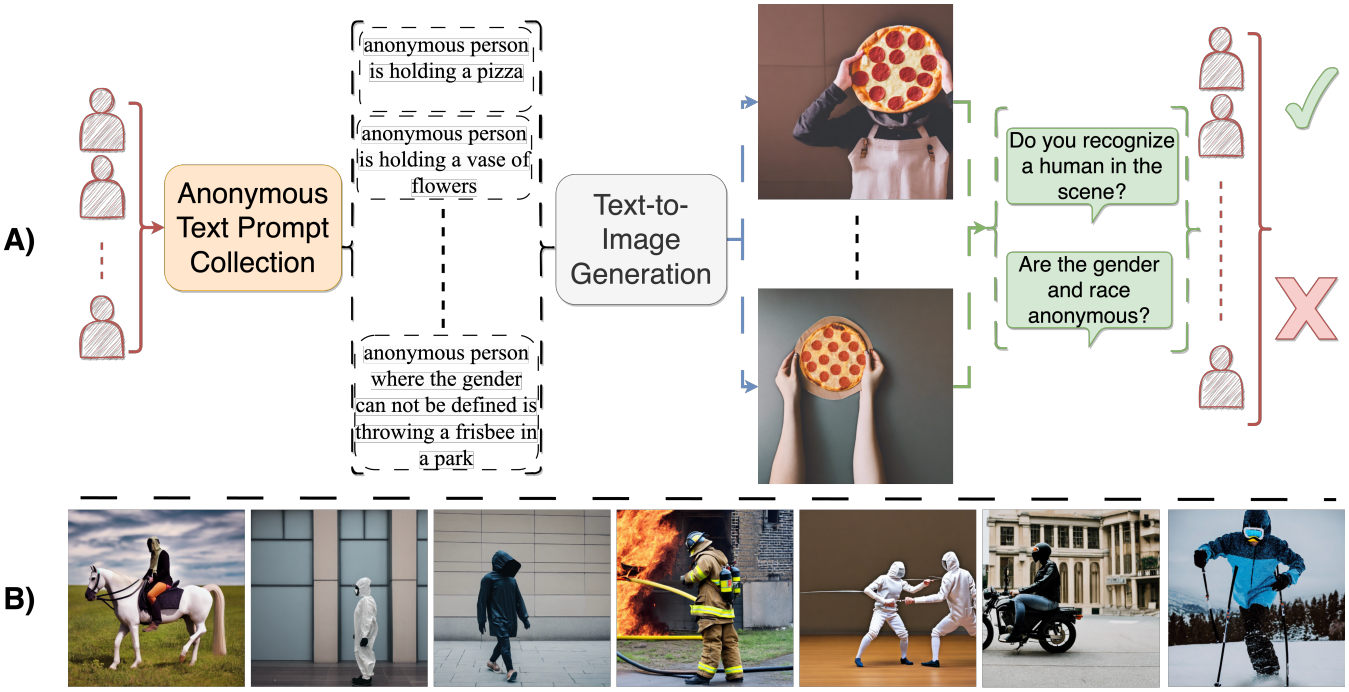

Figure 2: AnonymousBench. Part A is the detailed pipeline which demonstrates how our novel benchmark is collected. First annotators are asked to write a set of anonymous descriptions, then the generated prompts are fed to a text-to-image model to generate the images. Finally, as a verification step, the same annotators are asked again to filter out the gender or race recognizable images. Part B demonstrates some random samples of AnonymousBench.

and the model. In contrast, Women also Snowboard (Hendricks et al. 2018), GAIC (Tang et al. 2021), BA (Zhao et al. 2017), and DBA (Wang and Russakovsky 2021) do not utilize any additional learnable parameters to measure the bias. Women also Snowboard (Hendricks et al. 2018) defines the error rate as the number of gender misclassifications, i.e., whether the protected attribute words have been correctly predicted in the generated caption or not. Consequently, GAIC (Tang et al. 2021) formulates the gender bias as the difference in performance between the subgroups of a protected attribute, e.g., GAIC (Tang et al. 2021) creates three groups for gender: male, female, and not-specified. BA (Zhao et al. 2017) and DBA (Wang and Russakovsky 2021) calculate the co-occurrence between the desired bias attribute, e.g., gender, and the predicted caption.

Absolute-Bias Vs. Bias-Amplification. Another comparison criterion is whether the metric measures the magnitude of the bias introduced by the model over the bias already existing in the data. This type of metric can be interpreted as Bias-amplification metrics (Zhao et al. 2017) (Wang and Russakovsky 2021) (Hirota, Nakashima, and Garcia 2022) (Wang et al. 2019), where it answers the following question: "Does the model introduce extra bias than ground-truth dataset?" To this end, these models ground the model bias score to the data bias score. In contrast, (Zhao, Wang, and Russakovsky 2021) (Hendricks et al. 2018) (Tang et al. 2021) do not provide this valuable information; that's why we call them absolute-bias metrics.

Fairness Metrics Limitations

Table 1 summarizes the existing bias metric limitations.

Ignore Multi-Modality. Image captioning is a multi-modal task, making evaluating its bias and determining its source more challenging. However, to the best of our knowledge, non-of the existing image captioning bias metrics (Hendricks et al. 2018) (Tang et al. 2021) (Zhao, Wang, and Russakovsky 2021) (Zhao et al. 2017) (Wang and Russakovsky 2021) (Hirota, Nakashima, and Garcia 2022), as shown in Table 1, include the image while measuring the bias. All of them degrade the captioning task to a text generation task, which is a flawed approximation. Driven by this, we propose a novel metric termed ImageCaptioner^2, which respects the multiple modalities nature of the image captioning task by including both the image and the text while measuring the bias. In other words, we instead argue for a formulation of measuring the bias of generated caption from the image captioning model as a prompt-based image captioning task.

Limited context. Women also Snowboard (Hendricks et al. 2018), GAIC (Tang et al. 2021), BA (Zhao et al. 2017), and DBA (Wang and Russakovsky 2021) achieve a consistent performance as they use a fixed formula while computing the bias, e.g., co-occurrence, instead of using learnable classifiers, however, they measure the bias based on the appearance of the protected attributes in the caption and discard the rest of the caption in their bias formula. In contrast, our metric ImageCaptioner^2, considers the entire context while measuring the bias and determining its sources.

Inconsistency. A reliable evaluation metric should have the

Method	Women also snowbard	GAIC	Understanding racial biases	BA	DBA	LIC	Ours
Full-Context	✗	✗	✓	✗	✗	✓	✓
Multi-Modal	✗	✗	✗	✗	✗	✗	✓
Consistency	✓	✓	✓	✓	✓	✗	✓
Learnable	✗	✗	✓	✗	✗	✓	✓
Amplification Magnitude	✗	✗	✗	✓	✓	✓	✓

Table 1: Comparison of various bias metrics, i.e. Women also snowbard (Hendricks et al. 2018), GAIC (Tang et al. 2021), Understanding racial biases (Zhao, Wang, and Russakovsky 2021), BA (Zhao et al. 2017), DBA (Wang and Russakovsky 2021), LIC (Hirota, Nakashima, and Garcia 2022). Where Full-Context indicates, the whole caption is utilized, not only protected attributes. Multi-Modal reveals the metric incorporates the image alongside the text while measuring the bias of image captioning models. Consistency means the metric gives the same results on multiple runs. Learnable determines whether the metric exploits additional learnable parameters, e.g., language classifiers. Finally, Amplification-Magnitude demonstrates the metric provides information regarding the extra bias introduced by the model over the data.

following characteristics: 1) *Encapsulated*: could be applied to any method without assuming access to its internal parameters or weights. 2) *Consistent and reliable*: must give the same result and conclusion regardless of the multiple runs. We noticed an inconsistency in LIC (Hirota, Nakashima, and Garcia 2022), when varying the language encoders, where two language encoders are utilized as classifiers; BERT (Devlin et al. 2018), and LSTM (Hochreiter and Schmidhuber 1997). For instance, based on the results reported in the LIC paper (Hirota, Nakashima, and Garcia 2022), when the LSTM classifier is utilized, the LIC score indicates that the Transformer (Vaswani et al. 2017) is better than the UpDn (Anderson et al. 2018), where the LIC score is 8.7 and 9, respectively, as the lower LIC score indicates a better model. While based on the BERT classifier, the UpDn (Anderson et al. 2018) is much better than the Transformer (Vaswani et al. 2017), where the LIC score is 4.7 and 5.9, respectively. This motivates us to propose a more robust metric, termed ImageCaptioner^2, which, as shown in Table 1, belongs to the bias-amplification, learnable, and source-identification families.

Prompt-Based Bias Amplification Metric

Based on the aforementioned analysis and limitations, several questions and concerns arose:

1. Can we regard the multi-modality nature by incorporating the image while measuring the bias introduced by the image captioning models? To this end, we utilize the image captioning model to assess the bias introduced by the image captioning model, dubbed ImageCaptioner^2.
2. Can we design a bias metric that better matches the underlying pre-training model we assess? To this end, we design a formulation of measuring the bias of generated caption as prompt-based image captioning instead of using language classifiers (Hirota, Nakashima, and Garcia 2022).
3. Can we assess the inner bias of a model without introducing any additional parameters? To this end, we show an interesting property of our metric, called Self-Assessment, where we use the same image captioning model we need to assess to measure its own bias.
4. Does the existing scoring function match the objective of measuring the bias?

ImageCaptioner^2

Driven by the aforementioned limitations, we propose a prompt-based bias amplification metric, termed ImageCaptioner^2.

Bias-Amplification. We not only measure the severity of the bias in predicted captions but also detect the source of the bias. That means, on one hand, we measure the bias from the dataset; on the other hand, we estimate the bias of the caption model. As shown in Figure 1, two streams are utilized to evaluate an arbitrary image captioning model; 1) ground-truth stream (In green), 2) model stream (In blue). To measure the caption model bias B_m w.r.t the ground-truth bias B_d, a bias-amplification module is utilized, i.e., subtraction operation; $B_{amp} = B_m - B_d$. Other relational operator can be utilized, e.g., division, however the division could be thought of as a normalized version of the subtraction; $\frac{B_m}{B_d} \propto \frac{B_m - B_d}{B_d}$. The normalized version will make it hard to compare performances across different datasets, but provide similar meaning if the dataset is fixed.

Text Pre-processing. As shown in Figure 1, given an input image, the associated GT-caption, and the predicted caption, a text pre-processing module is utilized to refine the captions before feeding them to the prompt-based image captioners. The text pre-processing module performs two main functionalities: 1) Masking the protected-attribute indicative words. 2) Appending the prompt template to the caption. We hypothesize that if the fed captions are not biased, whether GT or predicted captions, then an arbitrary protected-attribute classifier performance should be around the random performance. Therefore, masking the protected-attribute indicative words is essential to make our hypothesis reasonable. For instance, if we are concerned by the gender bias, and the input caption is "A man sitting in front of his laptop computer", then the output of the masking operation should be "A [MASK] sitting in front of [MASK] laptop computer", where all the words that reveal gender information are replaced by mask token.

Image Matters. As shown in Table 1, all the existing metrics ignore the image stream while measuring the bias. This motivates us to propose a metric that respects the multi-modality nature by incorporating the input image. In other words, we design a formulation of measuring the bias in the caption as an image captioning task, not a plain text generation task.

Therefore, in contrast to the existing bias metrics, we are the first work that assesses the image captioning bias using image captioner instead of language classifiers.

Bias image masking. Consequently, masking the input image is as vital as masking the protected-attribute indicative words. To ensure our hypothesis is valid, we prevent all leakage sources, i.e., images and text. We assume access on bounding boxes or segmentation masks to mask the protected attribute visual clues. However, this assumption is practical as most of the existing image captioning models inherently contain an object detection phase, which we can utilize. Figure 1 demonstrates an example of a masked image.

Prompt Implementation

Motivation. One possible solution to adapt the image captioner to judge whether the model is biased is to stack a classification head. A key drawback of this solution is that the added classifier is trained for a different objective, i.e., classification objective, rather than the image captioning model objective, i.e., text generation. In contrast, we introduce a prompt-based metric to better match the underlying pre-training objective of the image captioner that we assess.

Prompt engineering. To this end, we reformulate the bias measurement by introducing a prompt function f_{prompt}, from a classification objective to a text generation objective by refining the input caption by adding a predefined template T. The prompt function f_{prompt} could be implemented in various ways, depending on the position of the empty slot, i.e., [Answer]. The cloze prompt function incorporates the empty slot in the middle of the caption, while the prefix prompt function incorporates it at the end. We designed our prompt function f_{prompt} in prefix fashion to fit in both language model families, i.e., autoregressive and masked language models (MLM). For instance, if the protected attribute a is gender, then the template T is incorporated to the masked input x_m, such as, $x_p = T?x_mA$, e.g., "What is the gender of the following sentence? A [MASK] sitting ... computer. [Answer]", or $x_p = x_mTA$, e.g., "A [MASK] sitting ... computer. Therefore the gender is [Answer]".

Implementation Details. During training, we fine-tune the prompt-based image captioner to predict the last word, i.e., [Answer]. More specifically, given the predicted caption x_m, the prompt T, and the masked image I_m, our prompt-based judge model predicts the answer A, e.g., male or female. Therefore the training objective is interpreted as follows:

$$\mathcal{L}^{CE} = log(P(A|x_mT, I_m)). \tag{1}$$

Then the bias scores, i.e., B_m and B_d, are interpreted as to what extent the model is confident while predicting the answer A;

$$B_{d,m} = \frac{1}{|\mathcal{D}|} \sum_{(x_m, I_m) \in \mathcal{D}} P(A|x_mT, I_m), \tag{2}$$

where D is the GT and predicted captions for B_d and B_m, respectively. During the inference, we deliberately ignore the model's predictions by injecting the input refined text as a predicted word at each time-step t, which could be interpreted as a teacher-forcing (Cho et al. 2014) with a different intention. As our main objective is to predict the [Answer] based on the input caption x_p, not the predicted one.

Method	Leakage ↓	LIC ↓	Ours ↓
NIC	-0.47	1.84	2.68
SAT	-0.47	1.96	1.64
FC	-0.61	1.56	6.29
Att2in	-0.51	3.07	6.17
UpDn	-0.65	1.24	6.64
Trans.	1.06	1.39	6.19
Oscar	-0.56	1.58	5.09
NIC+	-0.47	2.91	3.18
NIC+Equ.	-0.56	0.33	5.93

Table 2: Ablation study about different scoring functions; Eq. 4, across different models. i.e., NIC (Vinyals et al. 2015), SAT (Xu et al. 2015), FC (Rennie et al. 2017), Att2in (Rennie et al. 2017), UpDn (Anderson et al. 2018), Trans. (Vaswani et al. 2017), OSCAR (Li et al. 2020), NIC+ (Hendricks et al. 2018), NIC+Equ. (Hendricks et al. 2018). Leakage (Wang et al. 2019) exploits only the indicator function which measures the classification accuracy. LIC (Hirota, Nakashima, and Garcia 2022) considers also the confidence score alongside the classification accuracy. In contrast, we remove the accuracy measure, and define the bias score as only the confidence score, as shown in Eq. 4.

Self-Assessment

Self-Assessment could be interpreted as a special case of our metric; ImageCaptioner^2, where the same image captioning model, is used to measure its own bias. This allows us to fully take advantage of the parameters learned during the pre-training phase without adding any additional computations. Avoiding adding any new parameters does not only improve efficiency but also avoids adding an extra source of bias, making the evaluation more robust.

Confidence Is All You Need

Leakage (Wang et al. 2019) and LIC (Hirota, Nakashima, and Garcia 2022) estimate the gender bias for image classification and image captioning, respectively, using external classifiers f_{cls}, depicted in Eq. 3.

$$\mathbf{B_d} = \frac{1}{|\mathcal{D}|} \sum_{(y,a) \in \mathcal{D}} f_s(y, a). \tag{3}$$

Consider a sample (I, y, a), where I is the input image, y is the corresponding output, e.g., caption in image caption task, and a is the protected attribute, e.g., gender, both of them (Wang et al. 2019) (Hirota, Nakashima, and Garcia 2022) use a scoring function f_s that mixes up the accuracy objective with the bias objective. As shown in Eq. 4, the leakage (Wang et al. 2019) exploits only the indicator function which measures the classification accuracy. Consequently, LIC (Hirota, Nakashima, and Garcia 2022) argues that the uncertainty measure provides additional evidence for measuring bias, so LIC considers also the confidence score, as shown in Eq. 4.

Method	LIC ↓	Ratio ↓	Error ↓	BA ↓	DBA_G ↓	DBA_O ↓	Ours ↓
NIC	3.7	2.47	14.3	4.25	3.05	0.09	2.68
SAT	5.1	2.06	7.3	1.14	3.53	0.15	1.64
FC	8.6	2.07	10.1	4.01	3.85	0.28	6.29
Att2in	7.6	2.06	4.1	0.32	3.60	0.29	6.17
UpDn	9.0	2.15	3.7	2.78	3.61	0.28	6.64
Trans.	8.7	2.18	3.6	1.22	3.25	0.12	6.19
OSCAR	9.2	2.06	1.4	1.52	3.18	0.19	5.09
NIC+	7.2	2.89	12.9	6.07	2.08	0.17	3.18
NIC+Equ.	11.8	1.91	7.7	5.08	3.05	0.20	5.93

Table 3: The bias amplification results for the gender attribute on MS-COCO dataset using different models. i.e., NIC (Vinyals et al. 2015), SAT (Xu et al. 2015), FC (Rennie et al. 2017), Att2in (Rennie et al. 2017), UpDn (Anderson et al. 2018), Trans. (Vaswani et al. 2017), OSCAR (Li et al. 2020), NIC+ (Hendricks et al. 2018), NIC+Equ. (Hendricks et al. 2018). The ratio and the error are introduced in (Hendricks et al. 2018). The ratio is definded based on the number of sentences which belong to a female set to sentences which belong to a male set. The error rate is the number of gender misclassifications. BA (Zhao et al. 2017) and DBA (Wang and Russakovsky 2021) measure the bias based on the appearance of the protected attributes in the caption, i.e., co-occurrence.

Method	LIC ↓		Ours ↓	
	LSTM	BERT	SAT	GRIT
NIC	3.7 (1)	-0.8 (1)	2.68 (2)	1.02 (2)
SAT	5.1 (2)	0.3 (2)	1.64 (1)	0.61 (1)
FC	8.6 (5)	2.9 (5)	6.29 (8)	2.30 (5)
Att2in	7.6 (4)	1.1 (3)	6.17 (6)	2.78 (6)
UpDn	9.0 (7)	4.7 (6)	6.64 (9)	2.82 (7)
Trans.	8.7 (6)	5.9 (8)	6.19 (7)	2.90 (8)
OSCAR	9.2 (8)	4.9 (7)	5.09 (4)	2.21 (4)
NIC+	7.2 (3)	1.8 (4)	3.18 (3)	1.16 (3)
NIC+Equ.	11.8 (9)	7.3 (9)	5.93 (5)	3.08 (9)

Table 4: The bias amplification results for the gender attribute on MS-COCO datasets for LIC and our metric using different judge models across different image captioning models. i.e., NIC (Vinyals et al. 2015), SAT (Xu et al. 2015), FC (Rennie et al. 2017), Att2in (Rennie et al. 2017), UpDn (Anderson et al. 2018), Trans. (Vaswani et al. 2017), OSCAR (Li et al. 2020), NIC+ (Hendricks et al. 2018), NIC+Equ. (Hendricks et al. 2018). The down arrows indicates less is better. The ranking of captioning models is reported in red, which indicates to what extend the metric is consistent when changing the judging model.

$$f_s(y,a) = \begin{cases} \mathbb{1}[f_{cls}(y) = a] & Leakage \\ S_a(y) * \mathbb{1}[f_{cls}(y) = a] & LIC \\ S_a(y) & Ours \end{cases} . \quad (4)$$

In contrast, aligned with the bias objective, we remove the accuracy measure represented in the indicator function, and define the bias score as only the confidence score, i.e., $B_{d,m} = S_a(y)$ following Eq. 2. The assumption of such adjustment can be shown in the following case, taking the gender bias measurement as an example, where the protected attribute $a = female$. If the output confidence score $S_a(y)$ for male and female are 0.51 and 0.49, this means that the attribute classifier f_{cls} predicts *male*. According to the indicator function this given sample will be discarded, even if we know that the confidence score near 0.5 means there is no bias. Driven by this, we argue that confidence is all you need for bias scoring.

Experiments

Datasets and Models

We evaluate our proposed metric, ImageCaptioner[2], on two datasets, i.e., MS-COCO captioning dataset (Lin et al. 2014) for gender and race attributes, Artemis V1 (Achlioptas et al. 2021), and V2 (Mohamed et al. 2022) for emotions attribute. For gender and race, we validate the effectiveness of our metric on a wide range of image captioning models, i.e., NIC (Vinyals et al. 2015), SAT (Xu et al. 2015), FC (Rennie et al. 2017), Att2in (Rennie et al. 2017), UpDn (Anderson et al. 2018), Transformer (Vaswani et al. 2017), OSCAR (Li et al. 2020), NIC+ (Hendricks et al. 2018), and NIC+Eq (Hendricks et al. 2018). While for Artemis V1 (Achlioptas et al. 2021) and V2 (Mohamed et al. 2022), we explor SAT (Xu et al. 2015), and Emotion-Grounded SAT (EG-SAT) (Achlioptas et al. 2021) with its variants. The EG-SAT is an adapted version of SAT that incorporates the emotional signal into the speaker to generate controlled text. Two variants of EG-SAT are studied based on the source of the emotion: 1) Img2Emo. A pre-trained image-to-emotion classifier is utilized to predict the emotion. 2) Voting. Each image has, on average, eight different captions; therefore, the input emotion is conducted by a simple voting mechanism to pick the most frequent emotion in the GT captions.

Implementation Details

Masking Attribute Clues. To measure gender bias, we mask attribute-related information in both image and text. Specifically, for gender and racial bias, the image is masked using GT masks or predicted bboxes. In addition, the protected attribute-related words are replaced by a [MASK] token. In contrast, we do not mask any image region when focusing on emotion bias, as emotion attributes are not represented directly in artworks. Instead, we only mask sentimental words that leak the emotion attribute in captions.

Network and Training Configuration. We explore two different prompt-based image captioning models, i.e., SAT (Xu et al. 2015) and GRIT (Nguyen, Suganuma, and Okatani 2022) to act as classifiers. To train our prompt-based image captioner, we split the original validation set, around 10^4 images, into training, validation, and testing sets, i.e., 70%, 10%, and 20%, respectively. These splits are balanced based on the protected attribute. All captioning models are trained using the same training configurations mentioned in (Hirota, Nakashima, and Garcia 2022). The added template T is “Therefore, the gender is [Answer]”, "Therefore, The race is [Answer]" and “Therefore, the emotion is [Answer]” for

gender, race, and emotions, respectively. We train our prompt-based image captioner for 40 epochs from scratch, using the weight initialization strategy described in (He et al. 2015). Adam optimizer (Kingma and Ba 2014) and mini-batch size of 128 are used for training all our models.

Software and Hardware Details. Our metric is implemented in Python using the PyTorch framework. All experiments are conducted using four NVIDIA V100 GPUs.

Scoring Function Ablation Study

As shown in Table 2, the scoring function heavily influences the results. Whereas, the uncertainty or confidence score of the model is a direct reflection of the severity of bias. The higher confidence score of the model predictions indicates more severe bias. Leakage, LIC, and ImageCaptioner^2 share the same model weights, however, it is hard to get reasonable and insightful observations from the result of Leakage as most models achieve the same bias score; almost -0.5. Consequently, when the confidence score is added as in LIC, the bias scores influenced a lot, which indicates that confidence has the dominant role. This supports the importance of removing the accuracy indicator from the scoring function. Therefore, we adapt the scoring function to rely only on the confidence score, Eq. 3.

Gender Bias Benchmark on MS-COCO

We benchmark our metric across a wide range of captioning models, against the existing bias metrics. For fair comparison, we follow the training configurations mentioned in (Hirota, Nakashima, and Garcia 2022). Based on our metric, all captioning models amplify the bias, which is consistent with other bias metrics. In addition, we observe the same observation presented in (Hirota, Nakashima, and Garcia 2022) and (Wang and Russakovsky 2021), where the NIC-Equ amplifies the gender bias despite enhancing the classification accuracy. Interestingly, the equalizer almost tripled the bias score compared to NIC based on LIC. In contrast, the bias score is only doubled based on our measure. This decrease in the gap between NIC and NIC-Equ between LIC and our metric, 3x and 2x, respectively, is reasonable as the Equalizer allows models to make accurate predictions of the gender based on the visual region of the person. Therefore, intuitively, the equalizer effect will be included in our metric as it respects the multi-modality nature of the task.

Which Metric Is Better?

Despite the intuitive superiority of our method over the existing metrics, the ranking inconsistency among different bias metrics raises the importance of developing a method to evaluate the effectiveness of each bias metric. Consequently, this raises the importance of developing a method to evaluate the effectiveness of each bias metric quantitatively. Especially our metric and LIC. As both of them are learnable metrics, measure the amplification magnitude, and include the full-context while measuring the bias.

Human Evaluation. To fairly compare the different bias evaluation metrics, a human evaluation has to be conducted. However, it is hard design such an evaluation for the bias.

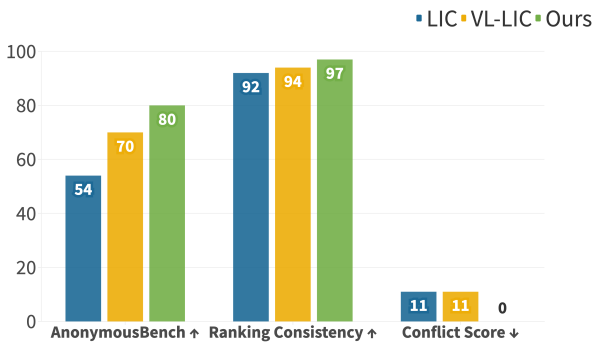

Figure 3: Comparison between our metric, LIC, and the multi-modal variant of LIC, termed VL-LIC, in terms of human evaluation using our AnonymousBench, ranking consistency, and conflict score.

For instance, one possible solution is asking the annotators to try to guess the gender given the AI generated captions, as shown in Figure 4. Unfortunately, the formulation will be not accurate enough as humans already biased, therefore this approach could lead to human bias measurement instead of metric evaluation. To tackle this critical point, we introduce AnonymousBench, depicted in Figure 2.

To prove the effectiveness of our method, we propose AnonymousBench, Figure 2; gender and race agnostic benchmark that consists of 1k anonymous images. First, we ask annotators to write 500 text prompts about various scenes with anonymous people. Secondly, Stable-Diffusion V2.1 (Rombach et al. 2022) is utilized to generate ten images per prompt resulting in 5k images. To address any potential bias (Zhang et al. 2023; Bakr et al. 2023) in the generated images, a human evaluation was conducted to filter out the non-agnostic images based on two simple questions; 1) Do you recognize a human in the scene? 2) If yes, Are the gender and race anonymous? Finally, the filtered images, 1k images, are fed to each model, to generate the corresponding captions. An unbiased model would predict gender-neutral words, e.g., person instead of man or woman, as the gender is not apparent. Therefore, the GT score is defined based on whether a human can guess the gender from the generated captions and averaged across the whole data. Table 5 demonstrates LIC, ImageCaptioner^2, and GT results on our proposed benchmark; AnonymousBench. The Pearson correlation is employed to measure the alignment between the metrics and human evaluation. As shown in Figure 3, our metric is more aligned with the human evaluation, where the correlation scores are 80% and 54% for our metric and LIC, respectively.

Learnable Metrics Consistency. The learnable metrics utilize additional language classifiers to measure the bias. Consequently, they may be inconsistent across different classifiers. To measure the consistency, LIC relies on the agreement between different classifiers on the best and the worst models in terms of bias. Following such a naive approach may lead to an inadequate conclusion. For instance, we observe an inconsistency in ranking, i.e., shown in red parenthesis in Table 4. In addition, there is no agreement on whether the

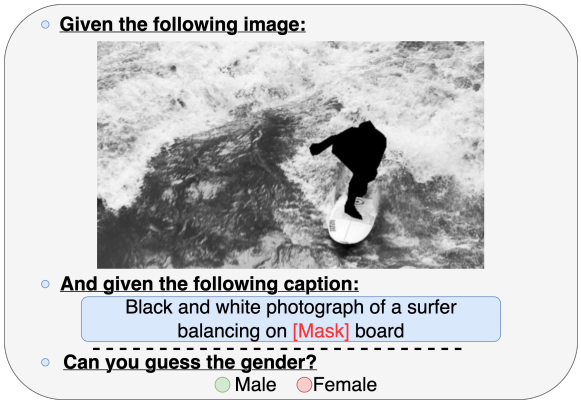

Figure 4: Human evaluation UI.

Method	LIC ↓	Ours ↓	GT ↓
NIC (Vinyals et al. 2015)	97.5	95.8	74.0
SAT (Xu et al. 2015)	92.0	94.0	33.0
FC (Rennie et al. 2017)	94.5	94.8	48.5
Att2in (Rennie et al. 2017)	93.0	92.3	48.5
UpDn (Anderson et al. 2018)	96.0	95.2	78.5
Trans. (Vaswani et al. 2017)	99.0	95.0	63.0
OSCAR (Li et al. 2020)	96.5	91.5	27.0
NIC+ (Hendricks et al. 2018)	97.5	95.6	87.5
NIC+Eq (Hendricks et al. 2018)	94.5	94.7	61.5

Table 5: Human evaluation results on AnonymousBench.

model is bias-amplified, e.g., NIC based on LSTM amplifies the bias (3.7) opposed to BERT (-0.8), where a positive score indicates the model amplifies the bias. Consequently, this motivates us to introduce two consistency measures to judge which learnable metric is classifier invariant; more consistent against classifier variations. 1) The conflict score (CS): Count the number of conflicts when changing the classifier. 2) Ranking consistency (RC): Measure the correlation between different classifiers.

Conflict Score. A reliable metric should be classifier invariant, at least in terms of the conclusion, i.e., whether the model is biased. Therefore, we introduce a conflict score to measure the percentage of mismatches between different classifiers, where the less conflict score is better. First, we classify the models into binary categories, biased and not biased, where positive means the model is biased. Then, we calculate the miss-classification rate among different classifiers. Based on results reported in Table 4 and Figure 3, our metric, ImageCaptioner^2, is more robust, as the conflict score is 0 and 11.11 % for ours and LIC, respectively.

Ranking Consistency. We utilize the Pearson correlation. Aligned with the conflict score, the correlation score, depicted in Figure 3 and Table 4, also probes that our metric is more robust than LIC, where the correlation score is 97% and 92% for ours and LIC, respectively. Where a higher correlation score indicates a more consistent metric.

Replacing GT Masks by Detectors

Relying on GT masks could be considered as a major drawback. Therefore, to show that our metric is orthogonal with the predicted masks, we replace GT masks, using two off-the-shelf detectors to predict the humans to mask them from the image. As shown in Table 6, both YOLOX (Ge et al. 2021) and MaskR-CNN (He et al. 2017) almost achieve the same bias score as the GT segmentation masks. In addition, they are fully correlated, which indicates that the detection phase does not introduce additional bias.

Method	GT-Masks	YOLOX	M-RCNN
NIC (Vinyals et al. 2015)	2.7	2.9	2.7
SAT (Xu et al. 2015)	1.6	1.4	1.8
FC (Rennie et al. 2017)	6.3	6.5	6.5
Att2in (Rennie et al. 2017)	6.2	6.3	6.2
UpDn (Anderson et al. 2018)	6.6	7.1	6.9
Trans. (Vaswani et al. 2017)	6.2	6.1	6.4
OSCAR (Li et al. 2020)	5.1	5.4	5.2
NIC+ (Hendricks et al. 2018)	3.2	3.3	3.6
NIC+Eq (Hendricks et al. 2018)	5.9	6.0	6.2

Table 6: Comparison of various detectors against the GT masks using ImageCaptioner^2.

Multi-Modal Classifier

To highlight the impact of incorporating the image while measuring the bias, we implement a simple vision-and-language classifier, termed VL-LIC, as a baseline. Accordingly, we have integrated ResNet50 as the visual backbone in conjunction with the BERT language encoder utilized by LIC. As depicted in Figure 3, incorporating the image, referred to as VL-LIC, demonstrates an enhancement in human correlation on our AnonymousBench and improved consistency compared to LIC, which relies solely on language. Furthermore, including the image lies in accessing a richer signal and capturing spurious correlations more accurately, as the generated caption alone may not provide comprehensive details encompassed by the image. However, the prompt-based metric remains superior due to its alignment with the underlying pre-trained model, which follows an auto-regressive approach by predicting subsequent words based on preceding ones. Furthermore, employing pre-trained vision and language classifiers as judge models introduces additional bias, thus posing challenges in drawing robust conclusions.

Conclusion

In this paper, we identify and address limitations in existing fairness metrics for image captioning, e.g., lack of consideration for the multi-modality nature of image captioning. We introduce ImageCaptioner^2, which incorporates the visiolinguistic aspects of image captioning to estimate bias more effectively. Leveraging a prompt-based image captioner, we reformulate the bias metric as a text generation objective aligned with the underlying pre-training objective. We evaluate our metric across 11 image captioning techniques on MS-COCO, Artemis V1, and Artemis V2 datasets, considering three protected attributes—gender, race, and emotions. Finally, our proposed AnonymousBench demonstrates superior alignment with human judgments, outperforming recent bias metric LIC by 80% and 54%, respectively.

References

Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; and Guibas, L. J. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11569–11579.

Alvi, M.; Zisserman, A.; and Nellåker, C. 2018. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Bakr, E. M.; Sun, P.; Shen, X.; Khan, F. F.; Li, L. E.; and Elhoseiny, M. 2023. HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models. *arXiv preprint arXiv:2304.05390*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cireşan, D. C.; Meier, U.; Masci, J.; Gambardella, L. M.; and Schmidhuber, J. 2011. High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*.

De Vries, T.; Misra, I.; Wang, C.; and Van der Maaten, L. 2019. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 52–59.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022. VOS: Learning What You Don't Know by Virtual Outlier Synthesis. *arXiv preprint arXiv:2202.01197*.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 771–787.

Hirota, Y.; Nakashima, Y.; and Garcia, N. 2022. Quantifying Societal Bias Amplification in Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13450–13459.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Khan, Z.; and Fu, Y. 2021. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*, 587–597.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The open images dataset v4. *International Journal of Computer Vision*, 128(7): 1956–1981.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137. Springer.

Liao, Y.; Xie, J.; and Geiger, A. 2022. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Mohamed, Y.; Khan, F. F.; Haydarov, K.; and Elhoseiny, M. 2022. It is Okay to Not Be Okay: Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21263–21272.

Nguyen, V.-Q.; Suganuma, M.; and Okatani, T. 2022. GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features. In *European Conference on Computer Vision*, 167–184. Springer.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7008–7024.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9: 1408–1424.

Stock, P.; and Cisse, M. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In

Proceedings of the European Conference on Computer Vision (ECCV), 498–512.

Tang, R.; Du, M.; Li, Y.; Liu, Z.; Zou, N.; and Hu, X. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, 633–645.

Thong, W.; and Snoek, C. G. 2021. Feature and label embedding spaces matter in addressing image classifier bias. *arXiv preprint arXiv:2110.14336*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, A.; Liu, A.; Zhang, R.; Kleiman, A.; Kim, L.; Zhao, D.; Shirai, I.; Narayanan, A.; and Russakovsky, O. 2022. REVISE: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 1–21.

Wang, A.; and Russakovsky, O. 2021. Directional bias amplification. In *International Conference on Machine Learning*, 10882–10893. PMLR.

Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.-W.; and Ordonez, V. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5310–5319.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.

Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; and Russakovsky, O. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 547–558.

Zhang, Y.; Jiang, L.; Turk, G.; and Yang, D. 2023. Auditing Gender Presentation Differences in Text-to-Image Models. *arXiv preprint arXiv:2302.03675*.

Zhao, D.; Wang, A.; and Russakovsky, O. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14830–14840.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.