Would Deep Generative Models Amplify Bias in Future Models?

Tianwei Chen[1*], Yusuke Hirota[1], Mayu Otani[2], Noa Garcia[1], Yuta Nakashima[1]
Osaka University[1], CyberAgent Inc.[2]
{chentw@is., y-hirota@is., noagarcia@, n-yuta@}ids.osaka-u.ac.jp, otani_mayu@cyberagent.co.jp

Abstract

*We investigate the impact of deep generative models on potential social biases in upcoming computer vision models. As the internet witnesses an increasing influx of AI-generated images, concerns arise regarding inherent biases that may accompany them, potentially leading to the dissemination of harmful content. This paper explores whether a detrimental feedback loop, resulting in bias amplification, would occur if generated images were used as the training data for future models. We conduct simulations by progressively substituting original images in COCO and CC3M datasets with images generated through Stable Diffusion. The modified datasets are used to train OpenCLIP and image captioning models, which we evaluate in terms of quality and bias. Contrary to expectations, our findings indicate that introducing generated images during training does not uniformly amplify bias. Instead, instances of bias mitigation across specific tasks are observed. We further explore the factors that may influence these phenomena, such as artifacts in image generation (*e.g.*, blurry faces) or pre-existing biases in the original datasets.*

1. Introduction

Emerging deep generative models, such as DALL-E 2 [39], Imagen [42], or Stable Diffusion [40], have shown remarkable capabilities in producing high-quality images. Trained on extensive datasets gathered from the internet [6, 44, 45, 48], these models can generate visually compelling images based on user-customized text inputs or prompts, sparking a surge of enthusiasm for image generation across the online community. However, concerns regarding social biases have been systematically identified [22], including gender bias [5, 8, 14, 29, 30, 46, 47, 49, 57, 59, 67], ethnicity bias [5, 8, 29, 34, 49], and geographical bias [4, 5, 34, 51]. In particular, previous work [5, 8, 30, 59] has highlighted the tendency of deep generative models to produce biased images even when prompted with ostensibly neutral inputs, uncovering unfair associations between specific social groups and certain attributes [29, 46, 57, 67]. A common example is the generation of images depicting occupations, such as doctors and nurses, which have been shown to be strongly tied to gender and race.

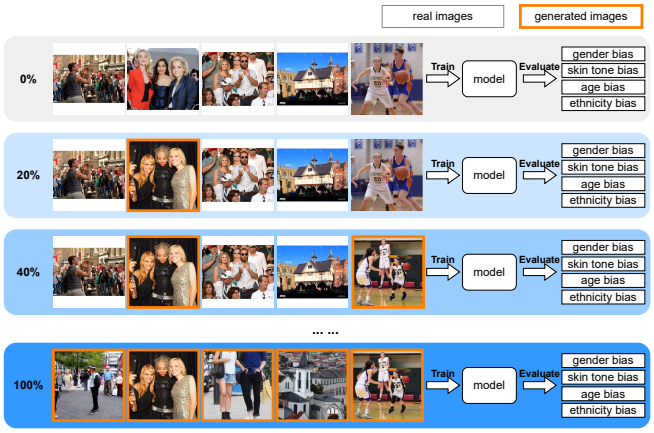


Figure 1. We investigate social biases in the training iterations of future models by simulating scenarios where generated images progressively replace real images in the training data.

Issues with bias tend to be attributed to the composition of the training data. Training images are frequently scraped from the internet with minimal efforts to filter out problematic samples and address representational disparities. Moreover, in the current context, generated images are continuously shared online and mixed with real images, which means that future computer vision models may inadvertently incorporate large portions of synthetically generated images into their training processes. Coupled with the increasing concerns about the presence of social bias in deep generative models, this raises the following question: *What consequences might arise if images generated by biased models become increasingly involved in the training process of future models?*

To address this question, we conduct experiments focusing on vision-and-language (VL) tasks within a scenario where generated images are progressively integrated into the training data. Specifically, we generate new images for

[*]Work done during internship at CyberAgent Inc.

COCO [27] and CC3M[1] [48] datasets using Stable Diffusion [40], and we gradually replace the original images in the datasets with their generated counterparts. Our evaluation covers four types of demographic bias – gender, ethnicity, age, and skin tone – across two tasks: image-text pre-training and image captioning. For image-text pre-training, we evaluate the bias introduced by OpenCLIP [7] on two downstream tasks, *i.e*. image retrieval [15, 68] and face attribute recognition [58]. For image captioning, we evaluate the performance of ClipCap [32] and Transformer [55] using bias metrics such as leakage (LIC) [20] and gender misprediction (Error) [18, 52].

Our experiments show that the behaviors of the evaluated biases are inconsistent and vary as we gradually replace original images with generated ones. In some cases, biases increase, while in others, they decrease. To understand this phenomenon further, we hypothesize two potential causes: 1) as existing datasets inherently contain biases [15, 68], if the bias introduced by the generated images aligns with the pre-existing biases in the dataset, it may not aggravate the existing bias, and 2) artifacts in Stable Diffusion's generations, particularly concerning the generation of human faces (*e.g*., blurred or poorly defined attributes), may lead models trained on such data to avoid learning demographic features. Overall, the key contributions of this paper are:

1. We show that, under our experimental setup, generated images from current deep generative models do not consistently amplify bias. Our experiments reveal different levels of bias for gender, ethnicity, age, and skin tone on both the COCO and CC3M datasets when increasing the number of generated images.
2. Through a set of follow-up experiments, we explore the underlying reasons behind these results, offering valuable insights into the dynamics between image generation models and existing datasets.
3. We propose recommendations for handling biased generated images in the training process of future models, contributing to the ongoing discourse on responsible and unbiased AI development.

While bias is not consistently amplified in our experiments, we find the presence of bias amplification in multiple instances concerning. Moreover, as our experiments are conducted on moderate-scale datasets with about 3 million images, representing about 130 times less data than the original CLIP [38], the impact of generated images on large-scale training remains uncertain. We believe that, as a community, addressing bias and ensuring models are safe for everyone should be a top priority. We hope our findings contribute to increased awareness of fairness in computer vision and inspire the creation of models with unbiased and equitable representations.

[1]CC3M is also known as Google Conceptual Captions or GCC.

2. Related work

Bias in pre-trained vision-and-language models Pre-trained VL models are not only used in downstream tasks through fine-tuning [24, 28, 65] but also in guiding model training [35, 40, 70] and serving as evaluation metrics [19, 61, 70]. With the proliferation of VL models, there is an increasing awareness about the inherent biases present in them [9, 15, 50, 58, 69]. For example, Wolfe *et al*. [58] evaluated the proximity of neutral text (*e.g*., "a photo of a person") and an attributive text (*e.g*., "a photo of a white person") in the CLIP embedding space [38]. The differences between demographic groups served as indicators of biases in the models. Chuang *et al*. [9] and Garcia *et al*. [15] explored performance gaps among demographic attributes (*e.g*., `man` and `woman` for gender, and `lighter` and `darker` for skin tone) in downstream tasks, such as classification and image retrieval. Overall, previous work [9, 15, 41, 58] has provided methodologies for detecting and evaluating bias in pre-trained VL models, especially in relation to gender and ethnicity. We leverage these approaches to anticipate potential bias in forthcoming datasets, particularly in scenarios where generated images dominate a significant portion of the online image sources, which is a plausible but underexplored scenario.

Synthetic data and pre-trained models Synthetically generated data is increasingly influencing the pre-training and fine-tuning processes of VL models, whether intentionally or unintentionally. On the one hand, synthetic data is used as an additional training resource when the original dataset is insufficient [60, 63, 66] or unreliable [54]. On the other hand, the widespread dissemination of synthetic images on the internet can inadvertently contaminate datasets [22]. Taori *et al*. [53] explored the data feedback loop and found that incorporating generated data into subsequent model training rounds could exacerbate dataset biases. Furthermore, Hataya *et al*. [17] showed that models trained on large portions of synthetic data dropped their performance. Building upon these insights, we study the repercussions of synthetic data on social bias in VL models.

3. Dataset contamination process

VL models are trained on pairs of images and text. The process for collecting this type of data typically begins with scraping the internet to gather a set of images $\mathcal{X} = \{x\}$, where x is an image. For smaller or moderately sized datasets [23, 27, 37], textual descriptions y for each image x are manually generated by crowdsourcing or in-house annotators, resulting in the set $\mathcal{Y} = \{y\}$. However, for large-scale datasets [6, 44, 45, 48], where generating specific annotations is unfeasible, text accompanying the images in the

original websites is used, often from the ALT[2] text. Subsequently, some form of filtering is applied to remove inappropriate content. Formally, let $p_{\mathcal{I}}(x)$ and $p_{\mathcal{T}}(y)$ represent the distributions of collected images and corresponding descriptions. All $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ can be seen as samples from $p_{\mathcal{I}}(x)$ and $p_{\mathcal{T}}(y)$, respectively. The textual description y is derived from $x \sim p_{\mathcal{I}}(x)$ through a framing process $y = f(x)$, which determines what aspects of x to describe.

Biases in the dataset-creation process are introduced from three main sources [12]. Firstly, biases are inherited from the original population of images on the internet,[3] in which content from specific demographic groups and geographical regions is overrepresented. Secondly, additional biases are introduced by the image descriptions provided by annotators or website authors, reflecting their stereotypes. Lastly, the filtering process itself can introduce additional bias; for instance, in the CC3M dataset, entities appearing less than 100 times were filtered out, potentially removing content from minority groups.

We define dataset contamination with generated images (hereafter referred to as *dataset contamination*) as a dataset wherein part of its population is replaced with generated images. That is, someone uploads to the internet images $x' = g(y')$ generated by a generative model g with a prompt y'. In this process, we operate under two assumptions: (1) a mental image $\bar{x}$ that people aim to achieve with a generative model also conforms to the distribution $p_{\mathcal{I}}(x)$, and (2) the image description process from the mental image $\bar{x}$ to a prompt y' has the same framing and bias as f. Given these assumptions, we infer that y' adheres to the distribution $p_{\mathcal{T}}(y)$ as $y' = f(\bar{x})$ and $\bar{x} \sim p_{\mathcal{I}}(x)$. Therefore, the distribution $p_{\mathcal{G}}(x)$ of generated images is given by:

$$p_{\mathcal{G}}(x) = \sum_{y} p_{\mathcal{T}\to\mathcal{G}}(x|y)p_{\mathcal{T}}(y), \tag{1}$$

where $p_{\mathcal{T}\to\mathcal{G}}(x|y)$ corresponds to the generative process $g(y)$. This means that we can generate images from descriptions $y \in \mathcal{Y}$ as described in [17]. Eventually, we create a dataset $\mathcal{D}(\alpha)$ by sampling images x with a prior α from:

$$\mathcal{D}(\alpha) = \{x \sim (1-\alpha)p_{\mathcal{I}}(x) + \alpha p_{\mathcal{G}}(x)\}. \tag{2}$$

This process of dataset contamination allows us to evaluate the impact of the generative model while keeping the other sources of bias consistent with the original dataset.

4. Bias evaluation tasks

The range of tasks in the scope of VL is extensive and diverse. For a survey, please refer to [31, 64]. In this work, we examine the effects of dataset contamination on two fundamental tasks: image-text pre-training and image captioning. Next, we outline bias evaluation in each of them.

4.1. Image-text pretraining

Image-text pertaining involves training a model to learn semantic correspondences between visual appearance and text, such as associating the word "rabbit" with and image of a rabbit. Models like CLIP [38] and its variants [7, 13, 25, 33, 62] are trained on large-scale image-text pairs sourced from the internet. CLIP-like models are reported to exhibit social biases, including gender [9, 15, 16, 41, 58], ethnicity [9, 15, 58], age [15, 58], and skin tone [15], and are susceptible to additional biases introduced by dataset contamination. We use OpenCLIP [7], an open-source variant, and assess its performance on text-to-image retrieval, self-similarity, and person preference.

Text-to-image retrieval Following Garcia *et al.* [15], where CLIP was shown to perform differently for different demographic attributes (*e.g.* images of men showed a higher recall at k (R@k) than images of women), we evaluate text-to-image retrieval performance. Text-to-image retrieval consists on finding the corresponding image given an input text. We compute R@k for different demographic attributes on PHASE [15] and COCO [27] datasets for OpenCLIP models trained on datasets $\mathcal{D}(\alpha)$.

Self-similarity Proposed by Wolfe *et al.* [58], *self-similarity* evaluates how images of an attribute group are distributed in the embedding space. The core idea is that if a CLIP-like model is trained on numerous images of a specific group with diverse descriptions in the contrastive training process, its encoders will attempt to distribute these images within a larger volume in the embedding space to differentiate them. Otherwise, images of an underrepresented group may occupy a smaller volume.

Formally, let $\mathcal{E}_a \subset \mathcal{E}$ denote the subset of the entire test set $\mathcal{E}$, containing only samples of a certain attribute group a. Self-similarity $\text{SS}(\mathcal{E}_a)$ for group a is given by:

$$\text{SS}(\mathcal{E}_a) = \frac{1}{|\mathcal{E}_a|^2 - |\mathcal{E}_a|} \sum_{x,x'} c(x, x'), \tag{3}$$

where $|\mathcal{E}_a|$ gives the number of samples in $\mathcal{E}_a$, $c(x, x')$ denotes the cosine similarity between x and x' in the embedding space,[4] and the summation is computed over all combinations of two samples x and x' in $\mathcal{E}_a$. A higher self-similarity means images in $\mathcal{E}_a$ are concentrated in the embedding space.

[2] ALT text refers to the text in the ALT attribute of HTML tags.

[3] If the scraping is random sampling, the population is identical to $p_{\mathcal{I}}(x)$, but typically this is not the case because of filtering.

[4] Letting e_{V} denote the CLIP visual encoder, $c(x, x')$ is defined as $c(x, x') = \cos(e_{\text{V}}(x), e_{\text{V}}(x'))$ where cos gives the cosine similarity.

Different treatments of attribute groups appear in the difference of $\mathrm{SS}(\mathcal{E}_a)$'s among a in attribute $\mathcal{A}$.[5] Self-similarity is defined over the learned embedding space, and the samples in that space give different distributions for different datasets; therefore, self-similarity cannot be compared across models. As we are interested in how broad the distribution for $a \in \mathcal{A}$ are in comparison with others in $\mathcal{A}$, we normalize self-similarity scores as:

$$\bar{\mathrm{SS}}(\mathcal{E}_a) = \frac{\mathrm{SS}(\mathcal{E}_a)}{\sum_{a \in \mathcal{A}} \mathrm{SS}(\mathcal{E}_a)/|\mathcal{E}_a|} - 1. \tag{4}$$

Person preference Another possible reflection of bias in the embedding space is whether a neutral description of an image represents images of a specific attribute group, *i.e.*, if a certain group is well-represented in a dataset, a neutral description may cover the attribute group. *Person preference* [58] evaluates this skew by comparing the similarities among a neutral description (*e.g.*, "a photo of a person"), a description with a specific attribute group (*e.g.*, "a photo of a white person"), and images of the group. Formally, let t_{N} and t_a denote the neutral description and one attributed by a. The person preference score over $\mathcal{E}_a$ is given by:

$$\mathrm{PP}(\mathcal{E}_a) = \frac{1}{|\mathcal{E}_a|} \sum_{x \in \mathcal{E}_a} \mathbb{1}[c(x, t_{\mathrm{N}}) > c(x, t_a)] \tag{5}$$

where $\mathbb{1}$ is the indicator function, and we abuse notation c to represent the cosine similarity between an image and a description, embedding them with appropriate encoders.

4.2. Image captioning

Image captioning is the task of generating descriptions for an input image. Descriptions generated by image captioning models [24, 32] have been found to reproduce bias, especially concerning gender and skin-tone [20, 52, 68]. We assess image captioning models trained on data contamination in terms on caption quality, LIC, and gender misprediction.

Caption quality Several automatic metrics have been proposed for evaluating captions quality, including BLEU [36], ROUGE [26], METEOR [3], CIDEr [56], and SPICE [1], which mainly involve a lexical comparison between the generated caption and the correspondent ground-truth caption. Alternatively, CLIPScore [19] evaluates the fidelity of a generated caption to the original image. In our experiments, we adopt BLEU-4, CIDEr, SPICE, and CLIPScore.

LIC To evaluate social bias amplification in image captioning models, Hirota *et al*. [20] proposed LIC. This metric evaluates whether the generated captions are more biased than the captions in the original trained dataset. For LIC, a set of captions is assumed to be biased if a protected attribute can be predicted without being explicitly mentioned. Specifically, an attribute classifier $h_a(y)$, which gives the likeliness of an attribute group a from a caption y, is trained on a training set $\mathcal{C}_{\mathrm{T}} = \{(y, a)\}$, where a is the ground-truth attribute group. All attribute-specific words[6] in the caption y are masked so that the prediction is not trivial. Then, given a validation set $\mathcal{C}_{\mathrm{V}}$, again with all attribute-specific words being masked, the model's leakage score is computed as:

$$\mathrm{LIC}_{\mathrm{M}} = \frac{1}{|\mathcal{C}_{\mathrm{E}}|} \sum_{(y,a) \in \mathcal{C}_{\mathrm{E}}} h_a(y) \mathbb{1}[\operatorname*{argmax}_{a'} h_{a'}(y) = a] \tag{6}$$

$\mathrm{LIC}_{\mathrm{M}}$ gives a higher value if the attribute group is correctly predicted with a higher confidence value even for the masked captions in $\mathcal{C}_{\mathrm{E}}$, suggesting that the attribute group can be easily predicted from captions.

The leakage score is also computed for the captions in the original dataset, *i.e.*, $\mathrm{LIC}_{\mathrm{D}}$ for $\mathcal{Y}$. The final amplification metric LIC is defined as the difference between the dataset and the model leakage as:

$$\mathrm{LIC} = \mathrm{LIC}_{\mathrm{M}} - \mathrm{LIC}_{\mathrm{D}}. \tag{7}$$

Gender misprediction Another bias evaluation metric for image captioning is the *Gender misspredict* or *Error* [18, 52], which measures gender mispredictions in the generated captions as:

$$\mathrm{Error} = \frac{N}{M}, \tag{8}$$

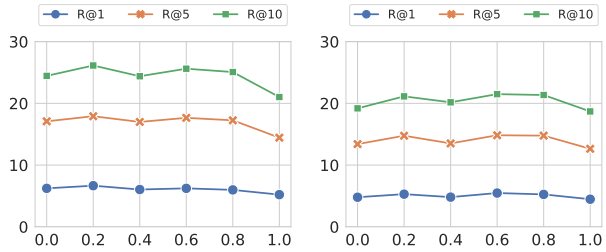
where M is the number of generated captions, and N is the number of captions among the M generated captions whose gender group is incorrectly predicted. Gender is considered incorrectly predicted if it contains any words in the attribute-specific word list for the gender opposite to the ground truth gender. For example, for the ground-truth group `man`, the gender in the generated caption is considered correct if there are no words from the `woman`-specific word list, such as *girl*.

5. Results on OpenCLIP

We train OpenCLIP [7] using various versions of the CC3M [48] dataset, each with different levels of dataset contamination. For dataset contamination, we use Stable Diffusion v1.5 [40] to generate images using the original captions as prompts. Due to the nature of the CC3M dataset, where images are provided as URL links and many of these links have expired, we are only able to retrieve $2,772,289$ valid images for our training data. Consequently, we generate images solely for the prompts corresponding to the

[5] For instance, the binarized gender attribute in PHASE [15] is given by $\mathcal{A} = \{\texttt{male}, \texttt{female}\}$.

[6] We use the same list of attribute-specific words as [20].



(a) COCO 2014 test set (b) Flickr30K test set

Figure 2. Image retrieval results on COCO 2014 test set and Flickr30k test set for different α. The performance of OpenCLIP remains consistent across different levels of dataset contamination.

available images. We randomly replace 20%, 40%, 60%, 80%, and 100% of original images with the images we generate, *i.e.* $\mathcal{D}(\alpha)$ for $\alpha = 0.0$ (the original CC3M dataset), 0.2, 0.4, 0.6, 0.8, and 1. Evaluation is conducted on five datasets, two for performance evaluation and three for bias evaluation. For performance evaluation, we use the COCO 2014 1K test set [27] and the Flickr30k test set [37]. For bias evaluation, we use the CC3M validation set using PHASE demographic annotations [15], the COCO validation set using gender and skin-tone annotations [68], and the whole FairFace dataset [21]. We run all experiments three times with different random seeds and report the average.

5.1. OpenCLIP performance

We first evaluate the performance of OpenCLIP trained under our experimental settings on two standard datasets: the COCO 2014 test set and the Flickr30K test set. We report text-to-image retrieval performance as R@k with $k = 1, 5, 10$. Results are shown in Figure 2, from which we observe that:

- Image retrieval results remain relatively constant for all levels of dataset contamination, from $\mathcal{D}(0.0)$ to $\mathcal{D}(1.0)$, in both datasets and for R@1, R@5, and R@10.
- Our reported results on OpenCLIP are considerably lower than those of the original CLIP. We attribute this difference to the disparity in the size of the training set. While our training is conducted with less than 3 million image-text pairs, the original CLIP model is trained on about 400 million samples.

In summary, the use of generated images for training OpenCLIP on the CC3M dataset appears to have minimal influence on the retrieval performance of its encoders. Next, we proceed to evaluate the impact of dataset contamination on the bias metrics.

5.2. Bias in OpenCLIP

As described in Section 4.1, text-to-image pertaining bias is evaluated on three metrics: text-to-image retrieval, self-similarity, and person preference. For text-to-image retrieval, we report results on the CC3M validation set with age, gender, skin-tone and ethnicity annotations from PHASE [15] (Figure 3) and the COCO validation set with gender and skin-tone annotations from [68] (Figure 4). For self-similarity and person preference, we report results on the FairFace dataset (Figures 5 and 6). From these results, we find the following trends with respect to bias:

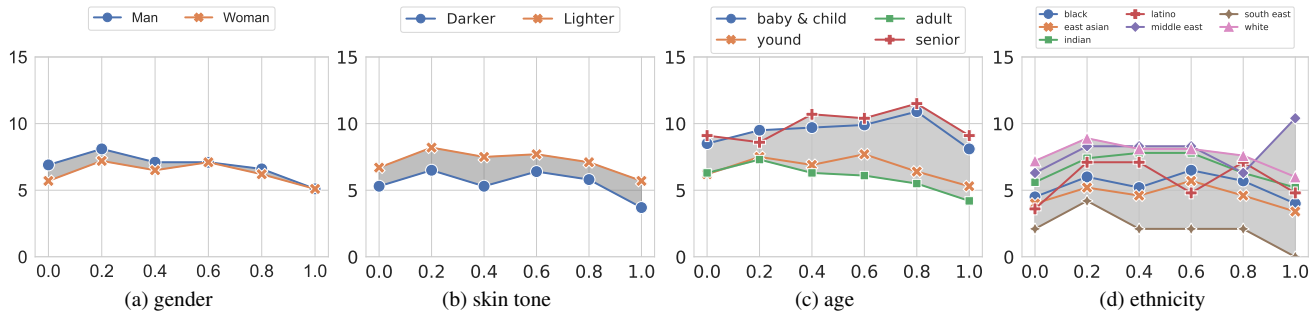
- **Consistent bias amplification**: We observe instances of consistent bias amplification, as illustrated in Figure 3c, where the text-to-image performance gap between the different age groups widens with increasing levels of dataset contamination.
- **Consistent bias mitigation**: In Figures 3a and 5a, we observe instances of consistent bias mitigation, where the gender gap is reduced for both text-to-image performance and self-similarity metrics. The gap in self-similarity for the age attribute is also consistently reduced, as shown in Figure 5b, indicating a bias mitigation effect with the increase of the dataset contamination parameter α.
- **Unaffected bias**: In some cases, bias remains unchanged. This is observed in Figure 3b, where the gap in text-to-image retrieval performance between lighter and darker-skin tone images remains constant for the different values of α from 0.0 to 1.0.
- **Ambiguous bias trends**: Across most instances, we do not discern a clear bias trend. In Figures 3a, 3d, 4b, 5c, 6a and 6c, we find no consistent pattern of bias changes, representing half of our experimental results. Unlike *unaffected bias*, the bias in these six experiments fluctuates, showing alternating increases and decreases. For instance, in Figure 6a, both the woman and man groups intermittently achieve the highest person preference scores. This suggests that multiple factors contribute to bias changes: some amplify bias, while others mitigate it, making bias changes unstable.

It is worth noting that the person preference scores show substantial variations in different experiments, surpassing 0.9 in gender and age (Figures 6a and 6b), while dropping to 0.2 for ethnicity (Figure 6c), despite the unclear trend of bias changes. This observation may be attributed to potential challenges associated with the generation of facial images with Stable Diffusion.

6. Results on image captioning

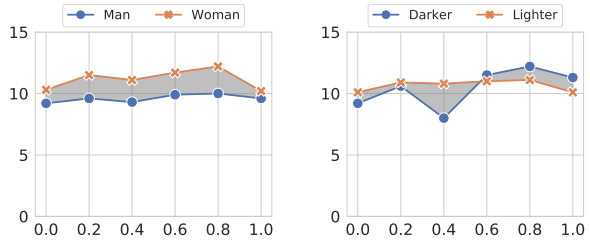
To analyze bias behavior in image captioning models trained with dataset contamination, we consider two models: Transformer[7] [55] and ClipCap [32]. Each model is trained on the COCO 2014 train set [27] with different levels of dataset contamination, ranging from $\mathcal{D}(0.0)$ to

[7]Transformer refers to a captioning model with a Transformer-based encoder-decoder where the encoder is ViT-B16 [11], and the decoder is BERT-base [10].



(a) gender (b) skin tone (c) age (d) ethnicity

Figure 3. $R@5$ on CC3M using PHASE annotations for different α. Bias is highlighted in gray as the difference between groups. We observe different trends: bias mitigation in Fig. 3a, consistency in Fig. 3b, amplification in Fig. 3c, and no clear trend in Fig. 3d.



(a) gender (b) skin tone

Figure 4. $R@5$ on COCO 2014 test set for different α. Bias is highlighted in gray as the difference between groups. Both gender and skin tone bias show ambiguous trends.

$\mathcal{D}(1.0)$. Evaluation is conducted in terms of caption quality and bias on the original COCO validation set using gender and skin-tone annotations from [68].

6.1. Image captioning performance

Image captioning results are presented in Table 1. Observing the image quality metrics (*i.e.*, BLEU-4, CIDEr, SPICE, and CLIPScore) we note the following:

- All lexical similarity-based metrics (*i.e.*, BLUE-4, CIDEr, and SPICE) either experience a gradual decrease or remain relatively stable from $\alpha = 0$, the original dataset, to 0.8. However, there's a significant drop between 0.8 and 1.0, suggesting that even a small amount of real images is necessary to maintain captioning performance.
- In contrast, the semantic similarity-based metric (*i.e.*, CLIPScore) remains unaffected by variations in dataset contamination, particularly evident in the case of the Transformer model. While ClipCap slightly improves in CLIPScore, we hypothesize that it is because of the use of CLIP in both image generation and image captioning processes. That is, Stable Diffusion uses CLIP to obtain the text embedding for a caption, so the generated image is strongly tied to it. Therefore, the training set $\mathcal{D}(\alpha)$ with larger α gives image-caption pairs that are close to each other in the CLIP embedding space. ClipCap trained with such a dataset thus only needs to learn the inverse process of the CLIP text encoder, *i.e.*, from an embedding to a caption, for these pairs, which can be easier than learning to fill the gap between images to captions. Thus, ClipCap may easily generate captions that match well with the corresponding images in the CLIP embedding space, consequently increasing CLIPScore.

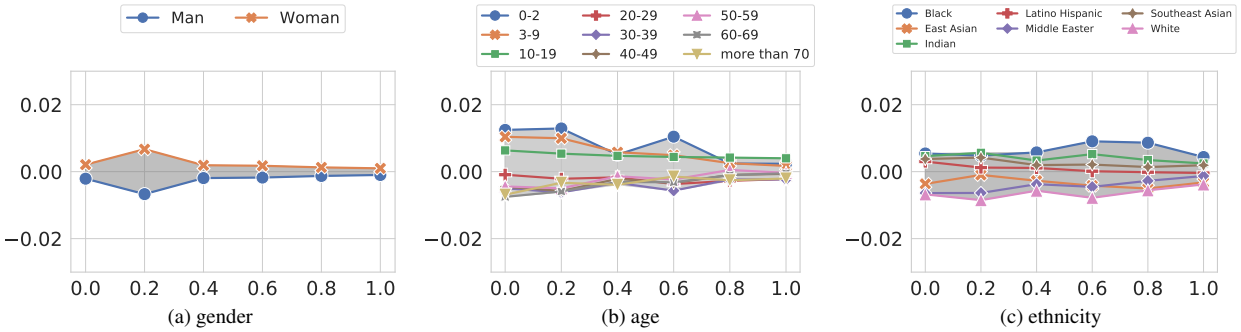
6.2. Bias metrics in image captioning

With regard to the bias metrics, which include LIC for gender (LIC-gender), LIC for skin-tone (LIC-skin), and gender mispredictions (error), the results are also presented in Table 1. We summarize our observations as follows:

- **No trend for gender bias**: LIC scores for gender show no noticeable trend across different values of α. In terms of gender mispredictions, similar to the LIC score, there is no clear tendency across the contamination ratios. Under our settings, we cannot draw any definitive conclusion about gender bias.
- **Skin-tone bias amplification**: While LIC for skin-tone on Transformer appears stable, on ClipCap it increases from 1.1 at $\alpha = 0$ to 3.1 at $\alpha = 1$. This trend could be attributed to Stable Diffusion accentuating the skin-tone bias present in the original dataset. For example, it has been found that, in the COCO dataset, indoor images tend to feature white people while black people tend to appear indoors [68]. Similar contextual biases have been observed in Stable Diffusion generations [5, 34].

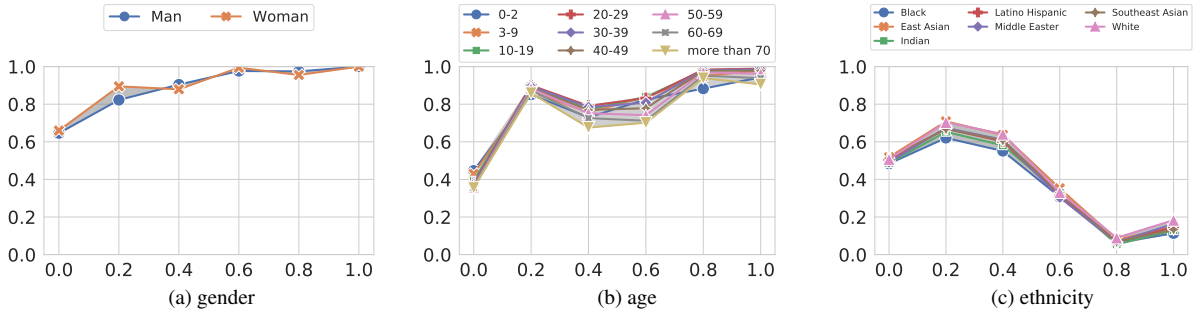
7. Analysis

Through our experiments, we observe the existence of different trends in the biases as we progressively replace real images with generated ones. To comprehend the underlying reasons behind this phenomenon, we explore potential factors based on our observations. We primarily focus on two possible explanations: (1) the inherent biases present within the original training datasets, and (2) the limitations of current deep generative models.



(a) gender (b) age (c) ethnicity

Figure 5. Self-similarity score of each group in the FairFace dataset for different α. Bias is highlighted in gray.



(a) gender (b) age (c) ethnicity

Figure 6. Person preference score of each group in the FairFace dataset for different α. Bias is highlighted in gray. None of the three figures show a clear tendency. Besides, the changes in bias are relatively small compared with the person preference scores.

Table 1. Captioning performance and bias metrics for ClipCap and Transformer.

	ClipCap							Transformer						
	Bias (↓)			Quality (↑)				Bias (↓)			Quality (↑)			
α	LIC-Gender	LIC-Skin	Error	BLEU-4	CIDEr	SPICE	CLIPScore	LIC-Gender	LIC-Skin	Error	BLEU-4	CIDEr	SPICE	CLIPScore
0	3.6	1.1	5.0	31.9	105.0	20.4	76.4	3.6	2.2	11.0	28.3	92.0	18.2	72.8
0.2	3.8	1.9	4.7	31.8	105.1	20.4	76.8	7.6	1.6	12.1	28.4	92.1	18.0	73.1
0.4	5.1	1.6	4.8	31.5	104.5	20.4	77.0	6.1	0.6	14.6	27.3	88.7	17.7	72.6
0.6	3.9	1.6	4.5	31.4	104.1	20.3	77.2	5.3	2.0	10.7	26.5	88.0	17.4	73.1
0.8	4.1	2.0	4.6	30.7	102.4	20.0	77.4	3.9	1.9	11.1	26.8	87.7	17.3	72.8
1.0	3.5	3.1	4.1	23.8	84.6	17.7	78.3	2.2	2.2	13.2	21.0	70.3	14.9	72.9

Inherent biases in original datasets Even though Stable Diffusion is known to produce biased images [5, 8, 29, 30, 34, 51, 57, 67], the original datasets, CC3M and COCO datasets, have also been found to be strongly unbalanced [15, 68]. For example, the CC3M validation set shows large gaps in perceived skin tone, with $3,166$ images of lighter v.s. 318 images of darker skin-tone people, and perceived ethnicity, with $2,231$ images of White people v.s. 16 images of Middle Eastern people [15]. Similarly, the COCO validation set, has been annotated with $7,466$ images of man v.s. $3,314$ images of woman and $9,873$ images of lighter v.s. $1,096$ images of darker skin-tone people [68]. If the disparities in representation within the original datasets resemble the biases in the images generated by Stable Diffusion, it is plausible that the biases remain unchanged as real images are progressively replaced with generated ones.

Failure of generation in Stable Diffusion Deep generative models like Stable Diffusion present several limitations beyond bias concerns. One prominent issue is the tendency for faces to become blurred when generating multi-

Figure 7. Blurry faces in the generated images. When this happens, the attributes (*e.g.*, gender and age) on the faces are hard to distinguish and further used in the model's training.

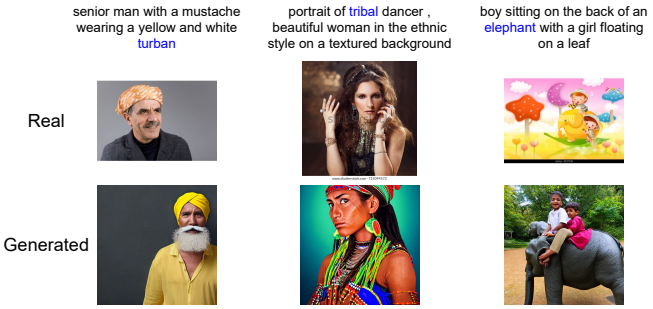


Figure 8. Stereotyping in the generated images. The words in blue may cause Stable Diffusion to generate stereotyped images.

ple people. Moreover, Stable Diffusion has been shown to stereotype certain culturally-associated words [51]. When examining the generated images in the training dataset, we find similar issues, as shown in Figures 7 and 8. These issues can impact bias: blurred faces may diminish gender or age biases, while stereotyping could potentially exacerbate ethnicity bias. This phenomenon could elucidate the gender bias mitigation observed in Figures 3a and 5a. Overall, due to the complexity of how bias originates and propagates across tasks, there is no one-size-fits-all solution to explain its causes and remedies.

8. Recommendations

From our experiments and analysis, we found that while images generated by Stable Diffusion exhibit bias across different demographic attributes, their use for training does not consistently amplify bias. This finding aligns with recent studies [2, 13, 43, 54] that use generated data from deep generative models for training. These studies highlight the diversity of effects that the generated data can have on model performance, potentially leading to performance improvements. Since the impact of generated data may depend on the original dataset and target task, we propose the following recommendations:

- **Bias-filtering preprocessing**: Considering the possibility that bias in the original dataset could be more pronounced than in deep generative models, we advocate for bias-filtering preprocessing during data collection from the internet, regardless of whether generated images are involved.
- **Caution with generation issues**: While generation issues like blurry faces may aid in bias mitigation in some tasks, they could potentially lead to bias amplification in others. Moreover, it is important not to regard generation issues as features, as they may be resolved in future iterations of generative models.

9. Limitations

- Due to the scale of current vision-and-language datasets like LAION-400M [44] and LAION-5B [45], our computational resources are insufficient for generating images and training models on such large datasets. Instead, our experiments are conducted using COCO and CC3M datasets, limiting the scope of insights to be drawn.
- The use of Stable Diffusion for image generation may overlook potential findings that could arise from other models with either more biased generations or better bias filtering capabilities.
- Our bias evaluation is focused on gender, age, ethnicity, and skin tone. The study does not explore all potential types of bias and leaves out the exploration of intersectional bias, leaving room for further investigation into additional dimensions of bias and fairness.

10. Conclusion

We investigated the impact of synthetic images generated by Stable Diffusion on bias in future models. We simulated a scenario where the generated images are progressively integrated into future datasets and evaluated bias in two downstream tasks: image-text pertaining with OpenCLIP and image captioning. Our findings revealed that the inclusion of generated images resulted in diverse effects on the downstream tasks, ranging from bias amplification to bias mitigation. Further visualization and analysis provided potential explanations underlying this phenomenon, including the inherent bias in the original datasets and the generation issues associated with Stable Diffusion.

11. Acknowledgment

This work is partly supported by JST CREST Grant No. JPMJCR20D3, JST FOREST Grant No. JPMJFR216O, JSPS KAKENHI Nos. JP22K12091 and JP23H00497.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 4

[2] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *ArXiv*, abs/2304.08466, 2023. 8

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, pages 65–72, 2005. 4

[4] Aparna Basu, R. Venkatesh Babu, and Danish Pruthi. Inspecting the geographical representativeness of images from text-to-image models. In *ICCV*, pages 5113–5124, 2023. 1

[5] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Y. Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. 2023. 1, 6, 7

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1, 2

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 2, 3, 4

[8] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. 2022. 1, 7

[9] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *ArXiv*, abs/2302.00070, 2023. 2, 3

[10] Devlin et al. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019. 5

[11] Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5

[12] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Comput. Vis. Image Underst.*, 223:103552, 2022. 3

[13] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. In *NeurIPS*, 2023. 3, 8

[14] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *ArXiv*, abs/2302.10893, 2023. 1

[15] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *CVPR*, pages 6957–6966, 2023. 2, 3, 4, 5, 7

[16] Melissa Hall, Laura Gustafson, Aaron B. Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. *ArXiv*, abs/2301.11100, 2023. 3

[17] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *ICCV*, 2023. 2, 3

[18] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, pages 771–787, 2018. 2, 4

[19] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 2, 4

[20] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, pages 13440–13449, 2022. 2, 4

[21] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, pages 1548–1558, 2021. 5

[22] Amelia Katirai, Noa García, Kazuki Ide, Yuta Nakashima, and Atsuo Kishimoto. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. *ArXiv*, abs/2311.18345, 2023. 1, 2

[23] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2016. 2

[24] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*, pages 121–137. Springer, 2020. 2, 4

[25] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 3

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, pages 74–81, 2004. 4

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 2, 3, 5

[28] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-Task Vision and Language Representation Learning. In *CVPR*, pages 10434–10443, 2020. 2

[29] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *ArXiv*, abs/2303.11408, 2023. 1, 7

[30] Abhishek Mandal, Susan Leavy, and Suzanne Little. Multimodal composite association score: Measuring gender bias in generative multimodal models. *ArXiv*, abs/2304.13855, 2023. 1, 7

[31] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language re-

search: A survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, 71:1183–1317, 2021. 3

[32] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 4, 5

[33] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. In *ECCV*, pages 529–544, 2022. 3

[34] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808, 2023. 1, 6, 7

[35] Jiefu Ou, Benno Krojer, and Daniel Fried. Pragmatic inference with a CLIP listener for contrastive captioning. In *ACL*, pages 1904–1917, 2023. 2

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 4

[37] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123:74–93, 2015. 2, 5

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022. 1, 2, 4

[41] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *NAACL-HLT*, pages 998–1008, 2021. 2, 3

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[43] Mert Bülent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR*, pages 8011–8021, 2023. 8

[44] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *ArXiv*, abs/2111.02114, 2021. 1, 2, 8

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 1, 2, 8

[46] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *ArXiv*, abs/2308.00755, 2023. 1

[47] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *ArXiv*, abs/2308.00755, 2023. 1

[48] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 1, 2, 4

[49] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan S. Kankanhalli. Finetuning text-to-image diffusion models for fairness. *ArXiv*, abs/2311.07604, 2023. 1

[50] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *ArXiv*, abs/2104.08666, 2021. 2

[51] Lukas Struppek, Dominik Hintersdorf, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *ArXiv*, abs/2209.08891, 2022. 1, 7, 8

[52] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *WWW*, pages 633–645, 2021. 2, 4

[53] Rohan Taori and Tatsunori Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *ICML*, pages 33883–33920, 2023. 2

[54] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *ArXiv*, abs/2306.00984, 2023. 2, 8

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2, 5

[56] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 4

[57] Jialu Wang, Xinyue Liu, Zonglin Di, Y. Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. *ArXiv*, abs/2306.00905, 2023. 1, 7

[58] Robert Wolfe and Aylin Caliskan. Markedness in visual semantic AI. In *FAccT*, pages 1269–1279, 2022. 2, 3, 4

[59] Yankun Wu, Yuta Nakashima, and Noa García. Stable diffusion exposed: Gender bias from prompt to image. *ArXiv*, abs/2312.03027, 2023. 1

[60] Yankun Wu, Yuta Nakashima, and Noa Garcia. Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In *ICMR*, pages 199–208, 2023. 2

[61] Dongsheng Xu, Wenye Zhao, Yi Cai, and Qingbao Huang. Zero-textcap: Zero-shot framework for text-based image captioning. In *ACM MM*, pages 4949–4957, 2023. 2

[62] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer.

Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18102–18112, 2022. 3

[63] David Junhao Zhang, Mutian Xu, Chuhui Xue, Wenqing Zhang, Xiaoguang Han, Song Bai, and Mike Zheng Shou. Free-atm: Exploring unsupervised learning on diffusion-generated images with free attention masks. *ArXiv*, abs/2308.06739, 2023. 2

[64] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *TPAMI*, 2024. 3

[65] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. VinVL: Revisiting Visual Representations in Vision-Language Models. In *CVPR*, pages 5575–5584, 2021. 2

[66] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kaixin Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. *ArXiv*, abs/2211.13976, 2022. 2

[67] Yanzhe Zhang, Lucy Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. *ArXiv*, abs/2302.03675, 2023. 1, 7

[68] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, pages 14810–14820, 2021. 2, 4, 5, 6, 7

[69] Kankan Zhou, Eason Lai, and Jing Jiang. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. In *AACL/IJCNLP*, pages 527–538, 2022. 2

[70] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *CVPR*, pages 17886–17896, 2022. 2