

# Resampled Datasets Are Not Enough: Mitigating Societal Bias Beyond Single Attributes

Anonymous ACL submission

## Abstract

We tackle societal bias in image-text datasets by removing spurious correlations between protected groups and image attributes. Traditional methods only target labeled attributes, ignoring biases from unlabeled ones. Using text-guided inpainting models, our approach ensures protected group independence from all attributes and mitigates inpainting biases through data filtering. Evaluations on multi-label image classification and image captioning tasks show our method effectively reduces bias without compromising performance across various models.

## 1 Introduction

Models trained on biased data can develop prediction rules based on spurious correlations (i.e., associations devoid of causal relationships), perpetuating and amplifying harmful stereotypes (Zhao et al., 2017). For example, image captioning models may generate gendered captions by associating gender with depicted activities (Zhao et al., 2023), location (Hendricks et al., 2018), or objects (Wang and Russakovsky, 2021). Dataset-level bias mitigation aims to reduce spurious correlations between labeled image attributes (e.g., teddy bear) and protected groups (e.g., woman). Resampling approaches balance the co-occurrence of each attribute with each group (Agarwal et al., 2022; Wang et al., 2020b). However, models can still exploit correlations between groups and sets of attributes (e.g., man with {dog, pizza, couch}), even when individual attributes are balanced (Zhao et al., 2023). Moreover, spurious correlations extend to unlabeled attributes, which current strategies do not address—e.g., gender disparities in image color statistics (Meister et al., 2023) or the person-to-object spatial distances (Wang et al., 2020a).

While equal group distributions in real-world datasets are challenging to achieve, generative text-to-image models now enable targeted image modifications (Rombach et al., 2022; Brooks et al., 2023;

Couairon et al., 2023). For example, bias detection methods alter image subjects’ appearance to assess counterfactual fairness (Joo and Kärkkäinen, 2020) or model bias (Smith et al., 2023; Brinkmann et al., 2023). However, manipulating individuals’ appearances without consent raises significant ethical and privacy concerns (Andrews et al., 2023; Yew and Xiang, 2022; Sobel, 2020; Ramaswamy et al., 2021a; Orekondy et al., 2018; Oh et al., 2016).

To address these challenges, we create training datasets with text-guided inpainting (Rombach et al., 2022), ensuring attribute distributions are independent of protected groups. Using masked person images and text prompts, we generate counterfactual images by inpainting only the masked regions, addressing ethical concerns of altering non-consensual images and ensuring equal representation of protected groups across attributes. We introduce data filters to mitigate biases from generative text-guided inpainting models (Bianchi et al., 2023; Cho et al., 2023; Bansal et al., 2022; Luccioni et al., 2023), evaluating images based on adherence to prompts, preservation of attributes and semantics, and color fidelity, validated by human evaluators. Unlike prior work (Wang et al., 2019, 2020b; Zhao et al., 2023; Agarwal et al., 2022), training on our counterfactual data decorrelates both labeled and unlabeled attributes from protected groups without impacting model performance. Comprehensive evaluations show our approach significantly reduces prediction rules based on spurious correlations in multi-label classification and image captioning across various architectures (e.g., ResNet-50 (He et al., 2016), Swin Transformer (Liu et al., 2021)), datasets (COCO (Lin et al., 2014), OpenImages (Krasin et al., 2017)), and protected groups (gender, skin tone). Our key contributions are summarized as follows:

- Introducing a framework for generating synthetic training datasets with group-independent image attribute distributions.



Figure 1: (a) Predicted objects by baseline ResNet-50 and with bias mitigation, i.e., over-sampling (Wang et al., 2020b) versus our method. (b) Generated captions by baseline ClipCap and with bias mitigation, i.e., LIBRA (Hirota et al., 2023) versus our method. Incorrect predictions, possibly affected by gender-object correlations, are in red.

- Proposing data filtering to mitigate biases introduced by generative inpainting models.
- Conducting quantitative experiments, demonstrating significant bias reduction in classification and captioning tasks compared to baselines.
- Identifying limitations of training on combined real and synthetic datasets, emphasizing the need for cautious synthetic data augmentation.

## 1.1 Related Work

Societal bias in datasets, characterized by demographic imbalances leading to spurious correlations, has been extensively studied (DeVries et al., 2019; Birhane et al., 2024; Birhane and Prabhu, 2021; Birhane et al., 2021; Wang et al., 2020a; Meister et al., 2023). These biases persist and can be exacerbated by multi-label classifiers (Zhao et al., 2017; de Vries et al., 2019; Wang et al., 2019) and image captioning models (Zhao et al., 2021; Hendricks et al., 2018; Hirota et al., 2022), disproportionately impacting historically marginalized groups such as women and individuals with darker skin tones (Garcia et al., 2023; Ross et al., 2020).

Two common approaches to bias mitigation are dataset-level and model-level. Dataset-level approaches leverage generative adversarial networks (GANs), counterfactual training dataset augmentation, and resampling. GANs create synthetic images to balance datasets and mitigate spurious correlations (Ramaswamy et al., 2021b; Sattigeri et al., 2019; Sharmanska et al., 2020), counterfactual data augmentation generates alternative scenarios to address biases (Kaushik et al., 2019; Wang and Culotta, 2021), and resampling bal-

ances the co-occurrence of attributes and protected groups (Agarwal et al., 2022; Wang et al., 2020b). Model-level approaches reduce bias through corpus-level constraints (Zhao et al., 2017), adversarial debiasing (Wang et al., 2019; Hendricks et al., 2018; Tang et al., 2021; Alvi et al., 2018), domain discriminative/independent training (Wang et al., 2020b), modified loss functions (Lin et al., 2017; Cui et al., 2019; Sagawa et al., 2019), and model output editing (Hirota et al., 2023). However, despite these advancements, existing mitigation methods focus on single labeled attributes, which can inadvertently increase models’ reliance on spurious correlations between protected groups and combinations of attributes (Zhao et al., 2023) or unlabeled attributes (Meister et al., 2023).

Recent progress in text-to-image generative models has enabled targeted image manipulation (Romach et al., 2022; Brooks et al., 2023; Couairon et al., 2023), which can help address bias in multimodal datasets. Nonetheless, these models have also been shown to perpetuate harmful stereotypes (Mandal et al., 2023; Zhang et al., 2023; Wang et al., 2023a; Struppek et al., 2022; Ungless et al., 2023; Naik and Nushi, 2023; Seshadri et al., 2023; Friedrich et al., 2023). In contrast to prior bias mitigation work, we use text-guided inpainting to generate synthetic training datasets that ensure equal representation of protected groups across all attribute combinations, whether labeled or unlabeled. To mitigate inpainting biases, we propose data filters, producing higher quality and less biased synthetic data. We go beyond previous work focused solely on gender bias mitigation (Joo and

Kärkkäinen, 2020; Smith et al., 2023; Brinkmann et al., 2023) by also addressing skin tone biases.

## 2 Method

We create training datasets with group-independent image attribute distributions by using masked person images and text prompts with an off-the-shelf diffusion model, as outlined in Figure 2.

### 2.1 Resampled Datasets Are Not Enough

We denote an image by  $x \in \mathcal{X}$ , a protected group by  $g \in \mathcal{G}$ , and an image attribute by  $a \in \mathcal{A}$ . A spurious correlation exists if  $p_{\mathcal{X}}(a | g) \neq p_{\mathcal{X}}(a)$ , indicating biases in the data. Resampling aims to remove these biases by adjusting the sampling process so that  $p_{\mathcal{X}}(a | g) = p_{\mathcal{X}}(a)$  for all  $g$  (de Vries et al., 2019; Wang et al., 2020b). This is done using a limited set of labeled attributes  $\mathcal{O} \subset \mathcal{A}$ , where attributes  $a$  are drawn from a distribution  $q(a)$  over  $\mathcal{O}$  and groups  $g$  are drawn from a uniform distribution  $u(g)$  over  $\mathcal{G}$  such that  $\mathcal{X}' = \{x \sim p_{\mathcal{X}}(x | g, a) | a \sim q(a), g \sim u(g)\}$ . This ensures  $p_{\mathcal{X}'}(a | g) = q(a)$  for  $a \in \mathcal{O}$  and  $g \in \mathcal{G}$ . However, this method has a limitation: it does not account for  $a$  being an unlabeled attribute or a combination of labeled and unlabeled attributes, making it difficult to sample  $x$  from  $p_{\mathcal{X}}(x | g, a)$  due to insufficient information about  $a$ . In short, while resampling can reduce biases, it is not always enough, especially when dealing with unlabeled or mixed attributes.

### 2.2 Text-Guided Inpainting

Suppose  $\mathcal{D} = \{(x_i, \omega_i, a_i, t_i^{(g)}) | 1 \leq i \leq n\}$  is a training set, where  $x \in \mathbb{R}^d$  is an image,  $\omega \in [0, 1]^d$  is a person mask,  $a$  is a labeled image attribute, a combination of labeled attributes, or an unlabeled attribute, and  $t^{(g)}$  is a text prompt containing a protected group-specific word  $g$ . To create a dataset with group-independent image attribute distributions, we utilize a text-guided inpainting model (Rombach et al., 2022). This model, guided by  $t^{(g)}$ , inpaints  $\omega$  in  $x$  with a synthetic person from protected group  $g$  described in  $t^{(g)}$ . For each tuple in  $\mathcal{D}$ , we generate  $m \in \mathbb{N}^+$  versions for each  $g \in \mathcal{G}$ , resulting in  $m \cdot |\mathcal{G}|$  samples:

$$\mathcal{D}_{\text{synthetic}} = \{(x_i^{(j,g')}, \omega_i, a_i, t_i^{(g')}) | 1 \leq i \leq n, g' \in \mathcal{G}, 1 \leq j \leq m\}, \quad (1)$$

where  $x_i^{(j,g')}$  denotes the  $j$ -th inpainted version of  $x_i \in \mathcal{X}$  for  $g'$  and  $t_i^{(g')}$  the modified text prompt where  $g$  in  $t_i^{(g)}$  is replaced with  $g'$ .

### 2.3 Societal Bias Data Filtering

Text-to-image generative models often perpetuate societal biases, portraying certain groups stereotypically, such as depicting women in brighter clothing (Bianchi et al., 2023; Cho et al., 2023; Bansal et al., 2022; Luccioni et al., 2023). Since these biases remain largely unaddressed (Smith et al., 2023; Brinkmann et al., 2023), we set  $m > 1$  in Equation (1) to generate multiple variations for each group. We propose filters to select the least biased inpainted images, evaluating images based on adherence to text prompts, preservation of attributes and semantics, and color fidelity. Specifically, for each tuple  $(i, g')$ , we select the highest quality and least biased version among the  $m$  versions to create a training dataset:

$$\mathcal{S}_{\text{synthetic}} = \{(x_i^{(j^*,g')}, \omega_i, a_i, t_i^{(g')}) \in \mathcal{D}_{\text{synthetic}} | \forall (i, g'), j^*\}, \quad (2)$$

where  $j^* = \arg \min_j \sum_k c_k \cdot r(s_k^{(i,j,g')})$ ,  $c_k \in \mathbb{R}$  are weights assigned to filters  $s_k$ ,  $s_k^{(i,j,g')}$  is the score obtained from applying filter  $s_k$  to image  $x_i^{(j,g')}$  for group  $g'$ , and  $r(s_k^{(i,j,g')})$  is the rank of the score for  $(i, g')$  in descending order, with lower ranks indicating less bias. Here,  $x_i^{(j^*,g')}$  is the selected inpainted image for tuple  $(i, g')$  that minimizes the sum of the ranks of the weighted filter scores, with  $j^*$  representing the index of the selected candidate image for tuple  $(i, g')$ .

Rather than creating an entire dataset of synthetic samples, we can augment  $\mathcal{D}$ :

$$\mathcal{S}_{\text{augment}} = \mathcal{D} \cup \{(x_i^{(j^*,g')}, \omega_i, a_i, t_i^{(g')}) \in \mathcal{D}_{\text{synthetic}} | \forall (i, g' \neq g), j^*\}. \quad (3)$$

The condition  $g' \neq g$  ensures that we only add inpainted images to  $\mathcal{D}$  for groups different from those originally present in  $x_i$ . In contrast to resampling,  $\mathcal{S}_{\text{synthetic}}$  and  $\mathcal{S}_{\text{augment}}$  ensure  $p_{\mathcal{X}'}(a | g) = p_{\mathcal{X}}(a)$  for all  $g \in \mathcal{G}$  without making assumptions about  $\mathcal{A}$ . Our proposed filters are introduced below.

**Prompt Adherence.** To evaluate the semantic alignment between  $x_i^{(j,g')}$  and  $t_i^{(g')}$ , we use CLIPScore (Hessel et al., 2021), which computes the cosine similarity between their CLIP embeddings (Radford et al., 2021). Formally,

$$s_{\text{prompt}}^{(i,j,g')} = \phi(x_i^{(j,g')}) \cdot \psi(t_i^{(g')}) \in [-1, 1], \quad (4)$$

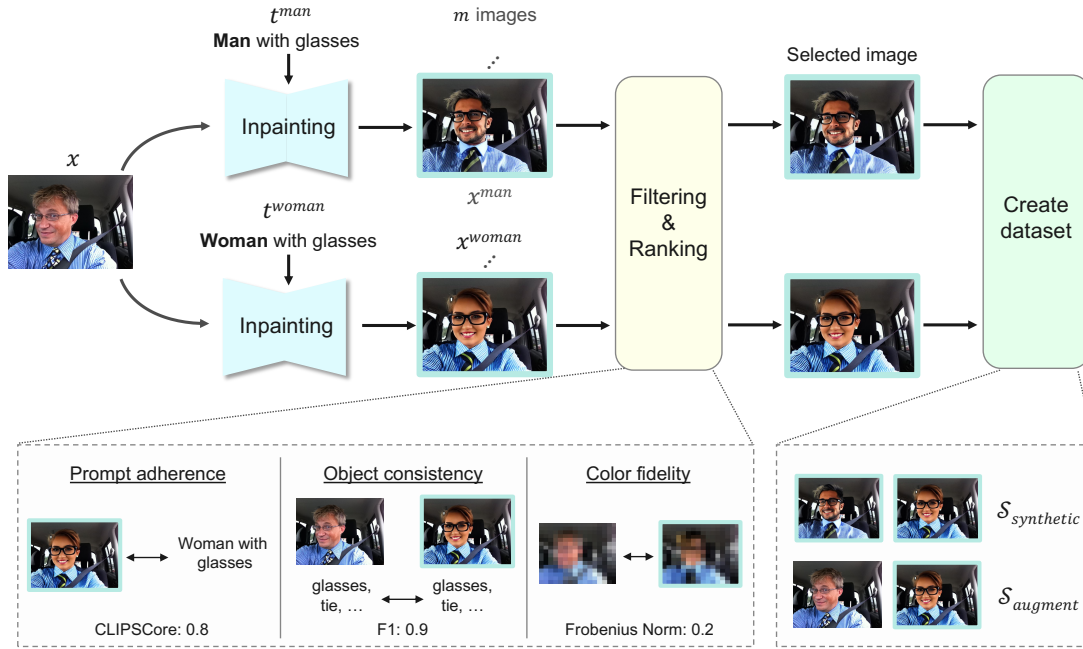


Figure 2: Overview of our pipeline for binary gender as a protected attribute. Original images are inpainted to synthesize diverse groups, maintaining consistent context. Synthesized images (highlighted in blue) are ranked using filters to select high-quality, unbiased samples (Module: Filtering & Ranking). Selected images are then used to construct datasets with group-independent image attribute distributions (Module: Create dataset).

where  $\phi$  and  $\psi$  are CLIP’s vision and text encoders, respectively. If  $s_{\text{prompt}}^{(i,j,g')} > s_{\text{prompt}}^{(i,j',g')}$ , then  $x_i^{(j,g')}$  better reflects the content described in  $t_i^{(g')}$ .

**Object Consistency.** To prevent the introduction of spurious correlations, such as generating objects not mentioned in  $t_i^{(g')}$  or reinforcing stereotypes (Bianchi et al., 2023; Cho et al., 2023; Bansal et al., 2022), we assess the object similarity between predicted objects in  $x_i^{(j,g')}$  and  $x_i$ . Concretely, we compute the F1 score (Sokolova et al., 2006) using a pretrained object detector (Zhou et al., 2022), denoted  $\eta$ :

$$s_{\text{object}}^{(i,j,g')} = \text{F1}[\eta(x_i^{(j,g')}), \eta(x_i)] \in [0, 1]. \quad (5)$$

If  $s_{\text{object}}^{(i,j,g')} > s_{\text{object}}^{(i,j',g')}$ , then  $x_i^{(j,g')}$  better preserves the integrity of the original unmasked scene in  $x_i$ .

**Color Fidelity.** Generative models can introduce subtler biases (Bansal et al., 2022; Bianchi et al., 2023), including those related to color (Meister et al., 2023). Addressing color biases is crucial as color choices can implicitly carry cultural or gendered connotations. To mitigate this, we down-sample  $x_i^{(j,g')}$  and  $x_i$  to  $14 \times 14$  pixels to focus on color rather than fine details, then measure the color difference using the Frobenius norm:

$$s_{\text{color}}^{(i,j,g')} = \|(x_i^{(j,g')})_{\downarrow 14 \times 14} - (x_i)_{\downarrow 14 \times 14}\|_{\text{F}}^{-1}. \quad (6)$$

If  $s_{\text{color}}^{(i,j,g')} > s_{\text{color}}^{(i,j',g')}$ , then  $x_i^{(j,g')}$  has better color fidelity to the original unmasked scene in  $x_i$ .

### 3 Experiments

Building on prior research (Zhao et al., 2017; Wang et al., 2019; Zhao et al., 2023; Hendricks et al., 2018; Zhao et al., 2021; Tang et al., 2021), we evaluate our synthetic dataset creation method on multi-label image classification and image captioning tasks using quantitative metrics, human studies, qualitative comparisons, and effectiveness analysis. Evaluations are conducted on test sets of real data.

**Implementation Details.** We inpaint the largest person in the image based on bounding box size, and if the second largest person exceeds 55,000 pixels, we also inpaint that region, using the person label for COCO. For image generation, we create  $m = 30$  inpainted images per group (e.g., woman, man) using guidance scales of 7.5, 9.5, and 15.0 to ensure diversity. Filter weights are set to 1 ( $c_k = 1$  for all  $k$ ), contributing equally. Results are based on five models trained with different random seeds. More details are in Appendices A and B.

#### 3.1 Multi-Label Classification

**Experimental Setup.** We focus on gender bias using the COCO dataset, retaining only images with gender-specific terms (e.g., woman, man) in

	ResNet-50			Swin-T			ConvNeXt-B		
	mAP	Ratio	Leakage	mAP	Ratio	Leakage	mAP	Ratio	Leakage
Original	<u>66.4</u>	6.3	13.4	<b>72.8</b>	4.0	14.3	<b>76.3</b>	4.6	18.2
Adversarial	63.3	—	<b>3.3</b>	67.8	—	<b>4.4</b>	69.6	—	<b>4.7</b>
DomDisc	57.4	4.1	15.4	65.4	4.6	16.8	68.8	4.5	19.1
DomInd	60.4	2.8	10.4	67.9	3.8	11.4	72.6	5.9	15.0
Upweight	64.9	9.1	8.3	71.5	6.3	9.8	75.0	5.6	12.9
Focal	66.1	6.3	12.0	<u>72.2</u>	3.8	13.3	<u>76.2</u>	3.8	16.2
CB	63.0	4.3	10.9	69.6	3.5	12.3	<u>73.8</u>	3.5	14.7
GroupDRO	64.1	3.0	11.4	70.8	<u>1.5</u>	12.6	75.3	4.2	16.4
Over-sampling	62.6	3.8	9.7	69.9	2.6	10.5	73.5	3.4	13.7
Sub-sampling	58.3	<u>2.0</u>	12.2	64.4	1.8	11.6	66.3	<u>2.2</u>	18.2
$\mathcal{S}_{\text{augment}}$ (Ours)	<b>66.9</b>	4.6	8.1	<b>72.8</b>	3.1	10.5	<b>76.3</b>	<u>2.2</u>	11.3
$\mathcal{S}_{\text{synthetic}}$ (Ours)	66.0	<b>1.1</b>	<u>7.5</u>	71.9	<b>1.4</b>	<u>8.4</u>	75.5	<b>1.2</b>	<u>8.2</u>

Table 1: Classification performance and gender bias scores of ResNet-50, Swin-T, and ConvNeXt-B backbones on COCO. Ratio is inapplicable to Adversarial due to its gender prediction module for mitigation. **Bold** and underline represent the best and second-best, respectively. For an unbiased model, Ratio = 1 and Leakage = 0.

	ClipCap				BLIP-2				Transformer			
	M	CS	Ratio	LIC	M	CS	Ratio	LIC	M	CS	Ratio	LIC
Original	<b>29.1</b>	<u>75.1</u>	2.5	2.2	<b>29.5</b>	75.1	5.7	4.7	<u>26.9</u>	<u>71.5</u>	4.7	4.7
LIBRA	28.9	74.9	6.5	<u>0.5</u>	29.0	<b>75.4</b>	6.3	<b>1.9</b>	<b>27.4</b>	<b>73.4</b>	6.7	2.3
Over-sampling	28.6	74.7	3.2	3.5	28.7	74.1	3.8	3.0	26.2	70.6	4.1	1.6
Sub-sampling	28.0	74.0	<u>1.4</u>	4.1	28.3	74.5	<u>1.4</u>	3.2	25.0	69.7	<u>2.0</u>	3.9
$\mathcal{S}_{\text{augment}}$ (Ours)	<u>29.0</u>	75.0	2.5	1.7	<u>29.4</u>	<u>75.3</u>	2.9	3.8	26.2	71.1	<u>2.6</u>	<u>1.5</u>
$\mathcal{S}_{\text{synthetic}}$ (Ours)	28.5	<b>75.3</b>	<b>1.3</b>	<b>0.3</b>	29.3	75.0	<b>1.2</b>	<u>2.5</u>	25.7	70.9	<b>1.4</b>	<b>0.5</b>

Table 2: Captioning quality and gender bias scores of ClipCap, BLIP-2, and Transformer backbones on COCO. M and CS denote METEOR and CLIPScore. **Bold** and underline represent the best and second-best, respectively. For an unbiased model, Ratio = 1 and LIC = 0.

289 their captions. This results in 28,487/13,487  
290 train/test samples. We focus on objects co-  
291 occurring with these terms, yielding 51 objects.  
292 ResNet50, Swin Transformer Tiny (Swin-T), and  
293 ConvNext models are fine-tuned using early stop-  
294 ping. Performance is assessed using mean average  
295 precision (mAP). Bias is quantified using leakage  
296 and ratio. Leakage measures how much the model’s  
297 predictions amplify the group’s information com-  
298 pared to the ground truth. A gender classifier  $f_g(y)$ ,  
299 predicting gender group  $g$  from input  $y$  (i.e., set of  
300 objects), is trained on a training set  $\mathcal{T} = \{(y, g)\}$ .  
301 For the test set  $\mathcal{T}'$ , the model’s leakage score is:

$$302 \text{LK}_M = \frac{1}{|\mathcal{T}'|} \sum_{(y,g) \in \mathcal{T}'} f_g(y) \mathbb{1} \left[ \arg \max_{g'} f_{g'}(y) = g \right] \quad (7)$$

303 The leakage score for the original dataset,  $\text{LK}_D$ ,  
304 is similarly computed. The final leakage is  
305 Leakage =  $\text{LK}_M - \text{LK}_D$ . Higher leakage indicates  
306 greater model exploitation of protected group infor-  
307 mation. Ratio measures the exploitation of attribute  
308 information for group prediction. By masking in-  
309 dividuals in test images and measuring the bias in

310 group predictions (e.g., #man-to-#woman ratio), de-  
311 viations from a ratio of 1 indicate attribute exploita-  
312 tion. We report Ratio =  $\max(r, r^{-1})$ , where  $r$   
313 is the observed ratio. This captures the magnitude of  
314 deviation from unbiased predictions consistently.

315 We compare our method with existing bias miti-  
316 gation techniques, including dataset-level meth-  
317 ods (Over-sampling (Wang et al., 2020b), Sub-  
318 sampling (Agarwal et al., 2022)) and model-level  
319 methods such as adversarial debiasing (Wang  
320 et al., 2019) (Adversarial), domain-independent  
321 training (Wang et al., 2020b) (DomInd), do-  
322 main discriminative training (Wang et al., 2020b)  
323 (DomDisc), loss upweighting (Byrd and Lipton,  
324 2019) (Upweight), focal loss (Lin et al., 2017) (Fo-  
325 cal), class-balanced loss (Cui et al., 2019) (CB),  
326 and group DRO (Sagawa et al., 2019) (GroupDRO).  
327 Additional results on the OpenImages dataset and  
328 skin tone bias mitigation are provided in Ap-  
329 pendix B.1, demonstrating consistent conclusions.

330 **Results.** Results are shown in Table 1. Our  
331 method,  $\mathcal{S}_{\text{synthetic}}$ , achieves the best balance by sig-  
332 nificantly improving both ratio and leakage while

maintaining a high mAP. Specifically,  $\mathcal{S}_{\text{synthetic}}$  achieves a near-ideal ratio of 1.1, low leakage of 7.5, and an mAP of 66.0 for ResNet-50, with similar trends observed for Swin-T and ConvNeXt-B.

Adversarial debiasing achieves lower leakage scores by removing gender information from intermediate representations. However, this method reduces mAP, indicating that object information may also be inadvertently removed. Over-sampling and sub-sampling methods address class imbalance but at the cost of model performance. Sub-sampling, in particular, reduces the ratio compared to over-sampling but results in worse mAP and increased leakage. This is likely due to the loss of diversity and information in the training data, which forces the model to rely more on the remaining features, increasing the influence of protected attributes.

In contrast,  $\mathcal{S}_{\text{synthetic}}$  generates diverse, high-quality synthetic samples, effectively balancing bias and variance. This approach avoids the pitfalls of other methods, resulting in superior performance metrics. While  $\mathcal{S}_{\text{augment}}$  performs similarly to the original dataset, it performs worse in terms of ratio and leakage compared to  $\mathcal{S}_{\text{synthetic}}$ .

### 3.2 Image Captioning

**Experimental Setup.** Using the COCO dataset (Section 3.1), we benchmark captioning models ClipCap, BLIP-2, and Transformer, which are fine-tuned using early stopping. Performance is evaluated with METEOR and CLIPScore. Bias is quantified using LIC and ratio, where LIC is a leakage-based metric that assesses the generation of group-stereotypical captions compared to ground-truth captions (i.e.,  $y$  is a caption in Equation (7)), and predicted group-related terms (e.g., woman) in captions used to compute ratio.

Bias mitigation baselines include dataset-level methods (over-sampling, sub-sampling) and the current state-of-the-art model-level method LIBRA (Hirota et al., 2023). LIBRA is a model-agnostic debiasing framework designed to mitigate bias amplification in image captioning by synthesizing gender-biased captions and training a debiasing caption generator to recover the original captions. Detailed results for skin tone bias mitigation, along with fine-tuning specifics, are provided in Appendix B.2, showcasing the generalizability of our approach.

**Results.** Results are shown in Table 2. Our method,  $\mathcal{S}_{\text{synthetic}}$ , significantly improves both ra-

tio and LIC while maintaining high METEOR and CLIPScore values. Specifically,  $\mathcal{S}_{\text{synthetic}}$  achieves a near-ideal ratio of 1.3, low LIC of 1.2, and a METEOR score of 29.3 for BLIP-2, with similar trends observed for ClipCap and Transformer.

While LIBRA effectively reduces LIC, it shows an increase in the ratio metric, indicating a trade-off between debiasing effectiveness and caption quality. Over-sampling and sub-sampling methods resulted in varying degrees of performance. Sub-sampling showed improved bias metrics compared to over-sampling but resulted in worse METEOR scores, especially for the Transformer model.

As in the multi-label classification task, we observe that although  $\mathcal{S}_{\text{augment}}$  significantly reduces bias compared to using the original dataset, there is a significant gap between it and  $\mathcal{S}_{\text{synthetic}}$  in terms of bias mitigation.

### 3.3 Analysis of Synthetic Artifacts

Recent studies show that text-to-image models introduce synthetic artifacts in images, which models may exploit (Qraitem et al., 2023; Corvi et al., 2023; Wang et al., 2023b). Our observations in Sections 3.1 and 3.2 suggest that bias persists with  $\mathcal{S}_{\text{augment}}$ , which augments the dataset with counterfactual images to balance group distributions. We hypothesize that  $\mathcal{S}_{\text{augment}}$  may lead to shortcut learning due to spurious correlations between minoritized groups and inpainted artifacts. In contrast,  $\mathcal{S}_{\text{synthetic}}$  distributes artifacts equally across all groups, avoiding this issue. To test this, we created a test set by inpainting random body parts using COCO-WholeBody annotations (Jin et al., 2020). Given an image, its caption, and body part annotations (e.g., left hand, right hand, head), we randomly selected a body part, created a mask using the Segment Anything Model (Kirillov et al., 2023), and performed inpainting with the caption as a prompt. We evaluated the consistency of ratios between the original and synthetic test sets; a gap indicates the exploitation of synthetic artifacts for gender prediction.

Table 3 presents scores for multi-label classification (ResNet-50, Swin-T) and image captioning (ClipCap, BLIP-2). The table includes the ratio of gender predictions (#man-to-#woman) for the original test set (Ratio<sub>orig</sub>) and the inpainted test set (Ratio<sub>inp</sub>), along with the relative difference ( $\Delta$ ) between these ratios. Results show a significant shift in gender predictions with  $\mathcal{S}_{\text{augment}}$ -trained models. Despite identical gender ratios in the original and





	Original	Inpainted
		
$\mathcal{S}_{\text{augment}}$	A man riding on the back of a motorcycle	A <b>woman</b> riding on the back of a motorcycle
$\mathcal{S}_{\text{synthetic}}$	A man riding on the back of a motorcycle	A man riding on the back of a motorcycle
		
$\mathcal{S}_{\text{augment}}$	Man flying in the air riding a motorcycle	<b>Woman</b> flying in the air riding a motorcycle
$\mathcal{S}_{\text{synthetic}}$	Man flying in the air riding a dirt bike	Man flying in the air riding a dirt bike

Figure 3: Predicted captions for the original (left) and unpainted (right) test images.

inpainted test sets (both set at 2.3), models trained with  $\mathcal{S}_{\text{augment}}$  predict woman much more frequently for the inpainted test set, indicated by the large relative differences. In contrast, models trained solely on synthetic data ( $\mathcal{S}_{\text{synthetic}}$ ) show minimal relative differences, indicating consistent gender predictions across original and inpainted test sets.

Figure 3 shows examples of synthetic images and predictions by ClipCap (trained on  $\mathcal{S}_{\text{augment}}$  or  $\mathcal{S}_{\text{synthetic}}$ ). The examples demonstrate inconsistent gender predictions with  $\mathcal{S}_{\text{augment}}$ ; specifically, the model tends to predict woman for inpainted test images, evidencing exploitation of synthetic artifacts.

### 3.4 Human Filter Evaluation

We conducted human evaluations on Amazon Mechanical Turk (Turk, 2012) to evaluate the effectiveness of our filters, aiming to determine if our filters prevent additional biases from inpainting models and ensure high-quality images. For 300 randomly selected original images, we analyzed inpainted images chosen by each filter combination. Evaluations focused on the similarity of 1) held/nearby objects, 2) object color, and 3) skin tone compared to the original images. Workers assessed differences between original and synthetic images for objects and their color, and selected skin tone classes using the Monk Skin Tone Scale (Schumann et al., 2023; Monk, 2023). Additionally, workers verified accu-

rate gender depiction through a sentence gap-filling exercise (e.g., “A \_\_\_\_ with a dog.”), where they must choose a protected group term to complete the sentence. More details are in Appendix B.3.

For the evaluation of the similarity of objects and their colors, scores are computed as the proportion of times the inpainted images are rated as similar. Regarding the skin tone and gender evaluations, the scores are calculated as the proportion of matching responses from workers between the original and inpainted images. All the scores range from 0 to 1. Table 4 summarizes the human evaluation and captioning performance of ClipCap trained on  $\mathcal{S}_{\text{synthetic}}$  (CS), with images selected by each filter. Notably, using all filters consistently received higher ratings across most criteria. In contrast, randomly selecting images without any filtering often leads to synthetic images differing significantly from the originals. This indicates that our filters are effective in mitigating additional biases introduced by the inpainting model. Furthermore, CLIPScore shows that using all filters improves captioning performance, highlighting its effectiveness in selecting higher-quality images.

### 3.5 Inherited Biases

To further discuss the potential biases introduced by the models used in our method, we conducted several assessments. First, for the object detector, we ran Detic (Zhou et al., 2022) on both real and synthetic images, achieving similar mAP scores of 32.0 for real images and 32.3 for synthetic images, indicating consistent performance. Second, addressing biases in CLIP, we acknowledge the potential biases inherent in the model. However, our use of object- and color-based filters helps mitigate these biases. Additionally, image classification and captioning results verify that our method effectively reduces gender and skin tone biases. Lastly, for the inpainting model, our filters effectively removed synthetic images that deviated from the prompt, altered color statistics, or introduced undescribed objects, as shown in Table 4. These assessments confirm that our method successfully mitigates biases without compromising performance.

### 3.6 Qualitative Results

We present qualitative examples of bias mitigation by applying our method ( $\mathcal{S}_{\text{synthetic}}$ ) in Figure 1. The results show that training models on  $\mathcal{S}_{\text{synthetic}}$  produces less biased outputs. For instance, in the classification task, the baseline ResNet-50 model

	ResNet-50			Swin-T			ClipCap			BLIP-2		
	Ratio <sub>orig</sub>	Ratio <sub>inp</sub>	$\Delta$	Ratio <sub>orig</sub>	Ratio <sub>inp</sub>	$\Delta$	Ratio <sub>orig</sub>	Ratio <sub>inp</sub>	$\Delta$	Ratio <sub>orig</sub>	Ratio <sub>inp</sub>	$\Delta$
Original	3.5	3.0	14.3	3.1	2.6	16.1	2.3	2.5	8.7	2.3	2.4	4.4
$\mathcal{S}_{\text{augment}}$	3.7	1.5	59.5	3.2	0.6	81.3	2.5	0.8	68.0	2.3	1.8	21.7
$\mathcal{S}_{\text{synthetic}}$	1.9	1.8	5.3	2.1	2.0	4.8	1.7	1.6	5.9	1.8	1.7	5.6

Table 3: Comparison of the original (Ratio<sub>orig</sub>) and inpainted (Ratio<sub>inp</sub>) versions of the COCO test set. The relative difference is denoted by  $\Delta = 100 \cdot \left| \frac{\text{Ratio}_{\text{orig}} - \text{Ratio}_{\text{inp}}}{\text{Ratio}_{\text{orig}}} \right| \%$ . A larger  $\Delta$  signifies a greater change.



Figure 4: Best/worst inpainted images for each filter in Section 2.3 and their combination (overall).

	Object	Color	Skin	Gender	CS
$s_{\text{prompt}} + s_{\text{object}} + s_{\text{color}}$	<b>0.57</b>	0.46	<u>0.29</u>	0.95	<b>75.3</b>
$s_{\text{prompt}} + s_{\text{object}}$	0.49	0.50	0.20	<b>0.99</b>	74.8
$s_{\text{prompt}} + s_{\text{color}}$	0.45	<b>0.56</b>	0.21	0.94	<u>75.2</u>
$s_{\text{object}} + s_{\text{color}}$	<u>0.53</u>	<u>0.52</u>	0.20	0.96	74.8
$s_{\text{prompt}}$	0.32	0.46	0.26	<u>0.97</u>	75.1
$s_{\text{object}}$	0.36	0.43	0.25	0.95	74.5
$s_{\text{color}}$	0.52	0.50	<b>0.30</b>	0.95	74.6
No filter	0.09	0.07	0.18	0.94	74.6

Table 4: Human evaluation and captioning quality (CLIPScore, CS in short) for each filter combination. Higher values indicate better alignment with original images. **Bold** and underline represent the best and second-best score for each metric.

and the over-sampling model incorrectly predict tie, due to its frequent co-occurrence with man in the training set. In contrast,  $\mathcal{S}_{\text{synthetic}}$  results in a gender bias-free prediction. Image captioning results further validate our approach. The baseline ClipCap model and LIBRA model generate the man-stereotypical word skateboard, whereas our method correctly predicts the object frisbee.

In Figure 4, we also present the best and worst inpainted images for each filter (prompt adherence, object consistency, and color fidelity), as well as their combination (overall). The results demonstrate each filter’s effectiveness, and combin-

ing them selects a high-quality image that closely resembles the original. For instance, the image judged worst by the object consistency filter lacks the object the man is holding, while the color fidelity filter’s worst image shows significant color changes in the man’s clothing. Combining these filters helps select an inpainted image that minimizes additional bias and closely matches the original.

## 4 Conclusion

We present a dataset-level bias mitigation pipeline that effectively reduces gender and skin tone biases by ensuring group-independent attribute distribution using synthetic-only images. Our findings indicate that mixing real and synthetic images introduces spurious correlations, underscoring the need for caution when augmenting datasets with synthetic data. Our work highlights the potential of synthetic data in bias mitigation and suggests further exploration into optimizing synthetic data generation and integration techniques for increased bias reduction.

## Limitations

**Binarized Group Classes and Intersectional Bias Analysis.** While acknowledging that gender and



549 skin tone exist on a spectrum, our data limitations  
550 necessitated a focus on binarized groups (i.e., man,  
551 woman), similar to prior work (Zhao et al., 2017;  
552 Wang et al., 2019; Zhao et al., 2023, 2021). Our  
553 analysis centered on gender and skin tone sepa-  
554 rately. However, our method can be extended to  
555 handle intersectional attributes (e.g., gender and  
556 skin tone) by inpainting with combinations of at-  
557 tributes (e.g., {woman, darker-skinned}, {woman,  
558 lighter-skinned}, {man, darker-skinned},  
559 {man, lighter-skinned}). We leave this exten-  
560 sion for future work to ensure a more compre-  
561 hensive and inclusive analysis of biases.

562 **Risks of Using Pre-trained Models.** As dis-  
563 cussed in Section 3.5, the pre-trained models em-  
564 ployed in our framework (e.g., inpainting model,  
565 object detector) may introduce inherent biases.  
566 While our analysis in Section 3.5 confirmed that  
567 these models do not adversely affect our method  
568 based on our evaluations, it is possible that some  
569 biases were not detected. Future research should  
570 focus on incorporating additional filters to further  
571 mitigate risks associated with pre-trained models.

572 **Residual Bias.** Our experimental results demon-  
573 strated that our method significantly mitigates soci-  
574 etal bias compared to existing methods. However,  
575 bias is not completely eliminated (e.g., leakage is  
576 not zero). Future work could explore further debi-  
577 asing by optimizing the weight of each filter (cur-  
578 rently, all filters are equally weighted), introduc-  
579 ing additional filters, and combining our method  
580 with existing bias mitigation techniques (e.g., focal  
581 loss).

582 **Extending to Additional Protected Groups.**  
583 Due to a lack of annotations for other protected  
584 attributes, our focus in this paper is on gender and  
585 skin tone biases. Nevertheless, our pipeline is ap-  
586 plicable to various protected attributes, such as age  
587 (e.g., “A woman with a dog” → “An elderly  
588 woman with a dog”). Future research should ex-  
589 plore the application of our method to additional  
590 protected attributes.

## 591 Ethics Statement

592 Our research involves the manipulation of image  
593 data to mitigate societal bias, raising important eth-  
594 ical considerations. We address these concerns by  
595 creating synthetic images that completely inpaint  
596 over identifiable individuals, thereby respecting pri-  
597 vacy and consent without altering their appearance.

598 Our approach aims to promote fairness and equity  
599 by ensuring diverse and unbiased representation  
600 in image datasets. We acknowledge the potential  
601 biases inherent in the pre-trained models used and  
602 have implemented filters to mitigate these biases  
603 as much as possible. Future work should continue  
604 to explore ethical guidelines and safeguards to en-  
605 sure the responsible use of generative models in  
606 research.

## References

- 608 Sharat Agarwal, Sumanyu Muku, Saket Anand, and  
609 Chetan Arora. 2022. Does data repair lead to fair  
610 models? curating contextually fair data to reduce  
611 model bias. In *WACV*.
- 612 Mohsan Alvi, Andrew Zisserman, and Christoffer Nel-  
613 låker. 2018. Turning a blind eye: Explicit removal  
614 of biases and variation from deep neural network  
615 embeddings. In *ECCV Workshops*.
- 616 Jerone Andrews, Dora Zhao, William Thong, Apostolos  
617 Modas, Orestis Papakyriakopoulos, and Alice Xiang.  
618 2023. [Ethical considerations for responsible data](#)  
619 [curation](#). In *Thirty-seventh Conference on Neural*  
620 *Information Processing Systems Datasets and Bench-*  
621 *marks Track*.
- 622 Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-  
623 Wei Chang. 2022. How well can text-to-image gen-  
624 erative models understand ethical natural language  
625 interventions? In *EMNLP*.
- 626 Federico Bianchi, Pratyusha Kalluri, Esin Durmus,  
627 Faisal Ladhak, Myra Cheng, Debora Nozza, Tat-  
628 sunori Hashimoto, Dan Jurafsky, James Zou, and  
629 Aylin Caliskan. 2023. Easily accessible text-to-  
630 image generation amplifies demographic stereotypes  
631 at large scale. In *FACCT*.
- 632 Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha  
633 Luccioni, et al. 2024. Into the laion’s den: Inves-  
634 tigating hate in multimodal datasets. *Advances in*  
635 *Neural Information Processing Systems Datasets and*  
636 *Benchmarks Track (NeurIPS D&B)*.
- 637 Abeba Birhane and Vinay Uday Prabhu. 2021. Large  
638 image datasets: A pyrrhic win for computer vision?  
639 In *WACV*.
- 640 Abeba Birhane, Vinay Uday Prabhu, and Emmanuel  
641 Kahembwe. 2021. Multimodal datasets: Misog-  
642 yny, pornography, and malignant stereotypes. *arXiv*  
643 *preprint arXiv:2110.01963*.
- 644 Jannik Brinkmann, Paul Swoboda, and Christian Bartelt.  
645 2023. A multidimensional analysis of social biases  
646 in vision transformers. In *ICCV*.
- 647 Tim Brooks, Aleksander Holynski, and Alexei A Efros.  
648 2023. Instructpix2pix: Learning to follow image  
649 editing instructions. In *CVPR*.

650	Jonathon Byrd and Zachary Lipton. 2019. What is the effect of importance weighting in deep learning? In <i>ICML</i> .	Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In <i>CVPR</i> .	702
651			703
652			704
653	Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In <i>ICCV</i> .	Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2023. Model-agnostic gender debiased image captioning. In <i>CVPR</i> .	705
654			706
655			707
656	Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. On the detection of synthetic images generated by diffusion models. In <i>ICASSP</i> .	Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. 2020. Whole-body human pose estimation in the wild. In <i>ECCV</i> .	708
657			709
658			710
659			711
660	Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2023. Diffedit: Diffusion-based semantic image editing with mask guidance. In <i>ICLR</i> .	Jungseock Joo and Kimmo Kärkkäinen. 2020. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In <i>International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia (FATE/MM)</i> .	712
661			713
662			714
663			715
664	Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In <i>CVPR</i> .	Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In <i>ICLR</i> .	717
665			718
666			719
667	Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone? In <i>CVPR Workshops</i> .	Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In <i>ICLR</i> .	720
668			721
669			
670	Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does object recognition work for everyone? In <i>CVPR Workshop on Fairness, Accountability Transparency, and Ethics in Computer Vision</i> .	Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. <i>arXiv preprint arXiv:2304.02643</i> .	722
671			723
672			724
673			725
674			726
675	Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In <i>ICLR</i> .	Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. <i>Dataset available from <a href="https://github.com/openimages">https://github.com/openimages</a></i> .	727
676			728
677			729
678			730
679			731
680			732
681	Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Lucioni, and Kristian Kersting. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. <i>arXiv preprint arXiv:2302.10893</i> .	Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. <i>arXiv preprint arXiv:2301.12597</i> .	734
682			735
683			736
684			737
685			
686	Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. 2023. Uncurated image-text datasets: Shedding light on demographic bias. In <i>Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition</i> , pages 6957–6966.	Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In <i>ICCV</i> .	738
687			739
688			740
689			
690			
691	Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In <i>ECCV</i> .	Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In <i>ECCV</i> .	741
692			742
693			743
694	Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In <i>ECCV</i> .	Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In <i>ICCV</i> .	745
695			746
696			747
697			748
698	Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In <i>EMNLP</i> .	Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In <i>CVPR</i> .	749
699			750
700			751
701		Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In <i>ICLR</i> .	752
			753

754	Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable bias: Analyzing societal representations in diffusion models. In <i>NeurIPS</i> .	Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In <i>CVPR</i> .	806
755			807
756			808
757			809
758	Abhishek Mandal, Susan Leavy, and Suzanne Little. 2023. Multimodal composite association score: Measuring gender bias in generative multimodal models. <i>arXiv preprint arXiv:2304.13855</i> .	Candace Ross, Boris Katz, and Andrei Barbu. 2020. Measuring social biases in grounded vision and language embeddings. <i>arXiv preprint arXiv:2002.08911</i> .	810
759			811
760			812
761			813
762	Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. 2023. Gender artifacts in visual datasets. In <i>ICCV</i> .	Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. <i>IJCV</i> .	814
763			815
764			816
765	Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In <i>CVPR</i> .	Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In <i>ICLR</i> .	818
766			819
767			820
768			821
769	Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. <i>arXiv preprint arXiv:2111.09734</i> .	Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. 2019. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. <i>IBM Journal of Research and Development</i> .	822
770			823
771			824
772	Ellis Monk. 2023. The monk skin tone scale.	Candice Schumann, Gbolahan O Olanubi, Auriel Wright, Ellis Monk Jr, Courtney Heldreth, and Susanna Ricco. 2023. Consensus and subjectivity of skin tone annotation for ml fairness. <i>arXiv preprint arXiv:2305.09073</i> .	825
773	Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. In <i>AIES</i> .		826
774			827
775	Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. 2016. Faceless person recognition: Privacy implications in social media. In <i>European Conference on Computer Vision (ECCV)</i> , pages 19–35. Springer.	Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. 2021. A step toward more inclusive people annotations for fairness. In <i>AIES</i> .	829
776			830
777			831
778			832
779			833
780	Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. 2018. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> , pages 8466–8475.	Preethi Seshadri, Sameer Singh, and Yanai Elazar. 2023. The bias amplification paradox in text-to-image generation. <i>arXiv preprint arXiv:2308.00755</i> .	834
781			835
782			836
783			837
784			838
785	Maan Qraitem, Kate Saenko, and Bryan A Plummer. 2023. From fake to real (ffr): A two-stage training pipeline for mitigating spurious correlations with synthetic data. <i>arXiv preprint arXiv:2308.04553</i> .	Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. 2020. Contrastive examples for addressing the tyranny of the majority. <i>arXiv preprint arXiv:2004.06524</i> .	839
786			840
787			841
788			842
789	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In <i>ICML</i> .	Brandon Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. 2023. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. <i>arXiv preprint arXiv:2305.15407</i> .	843
790			844
791			845
792			846
793			847
794	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. <i>OpenAI blog</i> .	Benjamin Sobel. 2020. A taxonomy of training data: Disentangling the mismatched rights, remedies, and rationales for restricting machine learning. <i>Artificial Intelligence and Intellectual Property (Reto Hilty, Jyh-An Lee, Kung-Chung Liu, eds.)</i> , Oxford University Press, Forthcoming.	848
795			849
796			850
797			851
798	Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. 2021a. Fair attribute classification through latent space de-biasing. In <i>IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)</i> .	Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In <i>Australasian joint conference on artificial intelligence</i> .	852
799			853
800			854
801			855
802			856
803	Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. 2021b. Fair attribute classification through latent space de-biasing. In <i>CVPR</i> .		857
804			858
805			859
			860

861	Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2022. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. <i>arXiv preprint arXiv:2209.08891</i> .	913
862		914
863		915
864		
865	Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating gender bias in captioning systems. In <i>WWW</i> .	916
866		917
867		918
868	Amazon Mechanical Turk. 2012. Amazon mechanical turk. Retrieved August.	
869		
870	Eddie L Ungless, Björn Ross, and Anne Lauscher. 2023. Stereotypes and smut: The (mis) representation of non-cisgender identities by text-to-image models. In <i>ACL</i> .	919
871		920
872		921
873		922
874	Angelina Wang, Arvind Narayanan, and Olga Russakovsky. 2020a. REVISE: A tool for measuring and mitigating bias in visual datasets. In <i>ECCV</i> .	
875		
876		
877	Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In <i>ICML</i> .	
878		
879	Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. 2023a. T2iat: Measuring valence and stereotypical biases in text-to-image generation. In <i>ACL</i> .	
880		
881		
882		
883	Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In <i>ICCV</i> .	
884		
885		
886		
887	Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020b. Towards fairness in visual recognition: Effective strategies for bias mitigation. In <i>CVPR</i> .	
888		
889		
890		
891		
892	Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In <i>AAAI</i> .	
893		
894		
895	Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023b. Dire for diffusion-generated image detection.	
896		
897		
898	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In <i>EMNLP: system demonstrations</i> .	
899		
900		
901		
902		
903		
904	Rui-Jie Yew and Alice Xiang. 2022. Regulating facial processing technologies: Tensions between legal and technical considerations in the application of illinois bipa. In <i>ACM Conference on Fairness, Accountability, and Transparency (FAccT)</i> , page 1017–1027.	
905		
906		
907		
908		
909	Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. 2023. Auditing gender presentation differences in text-to-image models. <i>arXiv preprint arXiv:2302.03675</i> .	
910		
911		
912		
	Dora Zhao, Jerone TA Andrews, and Alice Xiang. 2023. Men also do laundry: Multi-attribute bias amplification. In <i>ICML</i> .	913
		914
		915
	Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In <i>ICCV</i> .	916
		917
		918
	Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In <i>EMNLP</i> .	919
		920
		921
		922
	Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In <i>ECCV</i> .	923
		924
		925
		926
	<b>A Method Details</b>	927
	<b>A.1 Image Generation Settings</b>	928
	<b>Selection of People for Inpainting.</b> Following the previous works (Zhao et al., 2021; Misra et al., 2016), we apply inpainting to a person with the largest bounding box. In addition, if the second largest person’s box is larger than 55,000 pixels, the region is also inpainted. For COCO, we do this by using the person label and corresponding bounding boxes. For OpenImages, we use person-bounding boxes presented in More Inclusive Annotations for People (MIAP) annotations (Schumann et al., 2021), then we generate person masks within the boxes using Segment Anything Model (Kirillov et al., 2023).	929
		930
		931
		932
		933
		934
		935
		936
		937
		938
		939
		940
		941
	<b>Parameters of Image Generation.</b> In Section 2.2, we generate $m = 30$ inpainted images for each group (e.g., {woman, man} for binary gender). When generating the images, we use three different guidance scale parameters (7.5, 9.5, and 15.0) to generate diverse inpainted images (i.e., generating 10 images for each guidance scale). We use 6 NVIDIA A100-PCIE-40GB GPUs, resulting in a total of 72 hours to finish synthesizing images.	942
		943
		944
		945
		946
		947
		948
		949
		950
	<b>A.2 Visual examples of inpainted images &amp; failure cases</b>	951
		952
	We show the visual examples of the inpainted images after filtering in Figure 5 (for binary gender) and Figure 6 (for binary skin tone). The examples show that the inpainted images depict the target groups (e.g., woman and darker-skinned), keeping the rest fixed. In some cases, artifacts are noticeable, which enables us to identify synthetic images (e.g., the details of the faces are not clear), but they do not affect the downstream performance, as shown in the main paper.	953
		954
		955
		956
		957
		958
		959
		960
		961
		962



Figure 5: Examples of inpainted images for binary gender.

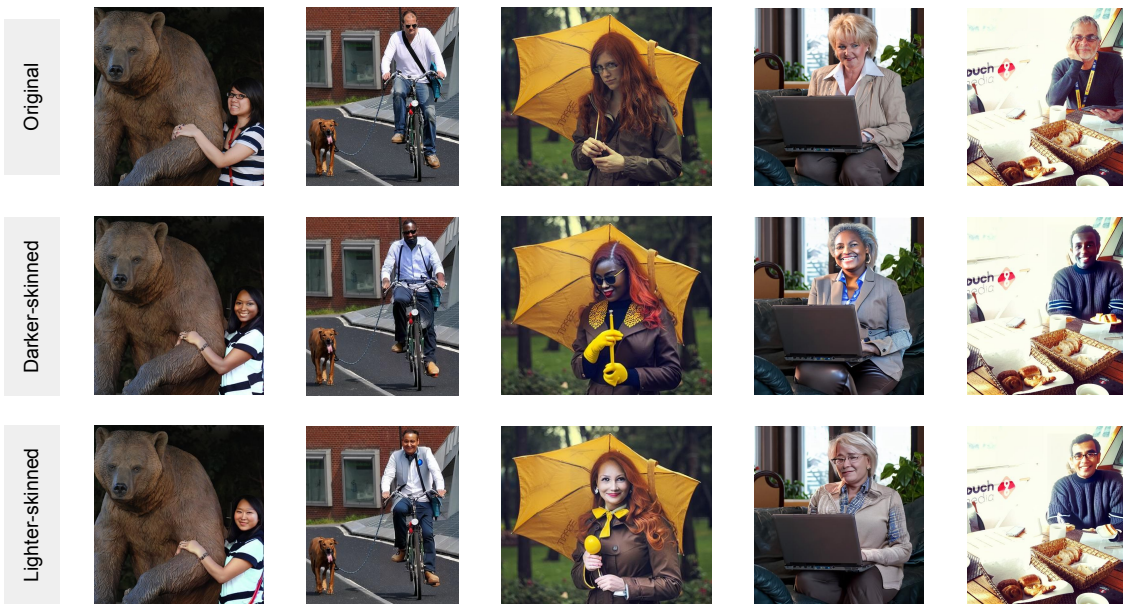


Figure 6: Examples of inpainted images for binary skin tone.

## B Experimental Settings and Additional Results

### B.1 Multi-Label Classification

**Datasets.** We use COCO (Lin et al., 2014) and OpenImages (Krasin et al., 2017). Following previous works (Zhao et al., 2017, 2023), we focus on attributes co-occurring with woman or man more than 100 times and remove person-related classes (e.g., person class), resulting in 51 and 126 attributes for COCO and OpenImages, respectively. The list of the attributes is as follows:

COCO: {sink, refrigerator, laptop, surfboard, vase, bottle, remote, donut, motorcycle, car, chair, suitcase, tv, knife, fork, couch, bus, toothbrush, bicycle, tie,

clock, microwave, teddy bear, frisbee, spoon, dog, truck, bench, backpack, skis, horse, sandwich, bed, handbag, umbrella, pizza, book, dining table, traffic light, banana, potted plant, tennis racket, cat, sports ball, kite, cake, wine glass, bowl, cup, oven, cell phone}.

OpenImages: {goggles, building, cloud, smile, tree, sunglasses, light, t-shirt, glasses, water, forehead, wall, sky, tire, roof, road, wheel, vehicle, land vehicle, car, tie, furniture, microphone, suit, clothing, fence, jeans, trousers, shirt, footwear, flooring, outerwear, coat, ceiling, floor, jacket, table, house, couch, mammal, hat, shoe, sports uniform, baseball (sport), cap,

978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997

	ResNet-50			Swin-T			ConvNeXt-B		
	mAP	Ratio	Leakage	mAP	Ratio	Leakage	mAP	Ratio	Leakage
Original	<u>42.3</u>	5.2	18.9	<u>45.3</u>	4.3	20.9	<u>46.0</u>	5.0	22.7
Adversarial	37.5	—	<b>8.3</b>	40.8	—	<b>11.3</b>	40.4	—	<b>12.3</b>
DomDisc	40.7	3.7	20.6	43.6	4.6	22.1	42.9	4.1	21.9
DomInd	40.3	3.7	19.1	42.7	3.5	20.2	43.4	2.6	22.0
Upweight	41.3	6.5	<u>13.1</u>	44.7	5.8	17.9	45.3	7.4	18.0
Focal	<b>43.0</b>	4.6	18.7	<b>45.4</b>	4.4	21.3	45.4	4.0	22.3
CB	40.5	5.2	18.0	42.6	3.9	19.8	43.9	4.6	21.5
GroupDRO	<u>42.3</u>	4.2	18.9	45.1	4.2	20.9	<b>46.1</b>	3.4	22.5
Over-sampling	38.5	3.3	15.0	41.1	4.0	16.1	41.7	5.2	18.4
Sub-sampling	38.3	<u>2.2</u>	18.3	41.2	<u>2.1</u>	19.8	39.8	2.8	21.7
$\mathcal{S}_{\text{augment}}$ (Ours)	42.0	1.9	16.0	44.9	2.4	18.0	45.5	2.6	19.0
$\mathcal{S}_{\text{synthetic}}$ (Ours)	41.4	<b>1.1</b>	14.6	44.4	<b>2.0</b>	17.6	44.7	<b>1.3</b>	<u>17.9</u>

Table 5: Classification performance and gender bias scores of ResNet-50, Swin-T, and ConvNeXt-B backbones on OpenImages. Ratio is inapplicable to Adversarial due to its gender prediction module for mitigation. **Bold** and underline represent the best and second-best, respectively. For an unbiased model, Ratio = 1 and Leakage = 0.

	ResNet-50		Swin-T		ConvNeXt-B	
	mAP	Leakage	mAP	Leakage	mAP	Leakage
Original	<b>65.8</b>	3.2	<b>72.2</b>	7.1	<b>75.9</b>	7.2
$\mathcal{S}_{\text{synthetic}}$ (Ours)	65.2	<b>2.3</b>	71.4	<b>3.7</b>	74.5	<b>5.9</b>

Table 6: Classification performance and skin tone bias scores of ResNet-50, Swin-T, and ConvNeXt-B backbones on COCO. **Bold** represents the best. For an unbiased model, Ratio = 1 and Leakage = 0.

baseball cap, bag, drawing, sun hat, musical instrument, baby, window, door, sweater, lake, chair, tableware, bottle, drink, handwriting, paper, food, tent, concert, drum, guitar, glove, sports equipment, blazer, art, painting, dress, flower, sneakers, screenshot, watercraft, beach, animal, grass family, plant, soil, desk, poster, bus, computer, personal computer, watch, mountain, helmet, bicycle helmet, bicycle wheel, bicycle, curtain, dance, football, ball (object), soccer, wedding dress, jewellery, bride, office building, laptop, toddler, shorts, hiking, fashion accessory, fedora, swimming, swimwear, camera, playground, weapon, ship, statue, boat, fast food, flag, soft drink, book, auto part, snow, carnivore, dog, horse, motorcycle, pole dance}.

**Training.** The models (ResNet-50 (He et al., 2016), Swin-T (Liu et al., 2021), and ConvNeXt-Base (Liu et al., 2022)) are initialized with ImageNet (Russakovsky et al., 2015) pre-training, and fine-tuned with early stopping using a validation set split from the training set (20% of the training set). The optimizer is Adam (Kingma and Ba, 2015),

batch size is 32, and a learning rate is  $1 \times 10^{-5}$ . For binary gender, the classification layers predict both protected groups (i.e., {woman, man}) and object classes. For binary skin tone, the models only predict object classes as ground-truth skin tone labels are not available.

**Results for OpenImages.** We show the complete results of the experiments in the main paper: gender bias on OpenImages (Table 5). The results show that all the insights described in the main paper are consistent across the datasets.

**Results for skin tone bias.** Previous bias mitigation methods face a significant limitation, requiring protected group labels for all training set samples (Zhao et al., 2017; Wang et al., 2019; Agarwal et al., 2022). They typically focus on gender as a protected attribute due to its prevalence in captions (Misra et al., 2016), allowing for label inference through gender-related terms. In contrast,  $\mathcal{S}_{\text{synthetic}}$  applies to attributes without labels, such as skin tone. We use our pipeline (excluding the color fidelity filter, as we aim to modify skin tone) on binary skin tone categories (i.e.,  $\mathcal{G} = \{\text{darker-skinned, lighter-skinned}\}$ ) using COCO. We evaluate skin tone bias using *leakage* only since *ratio* requires models to predict pro-

1046 tected groups, and there are no skin tone annota-  
1047 tions for the COCO training set. Results are shown  
1048 in Table 6, demonstrating consistent conclusions  
1049 with gender bias.

## 1050 B.2 Image Captioning

1051 **Training.** We benchmark three captioning mod-  
1052 els: ClipCap (Mokady et al., 2021), BLIP-2  
1053 (Li et al., 2023), and Transformer (i.e., the  
1054 Transformer-based encoder-decoder model com-  
1055 posed of Vision Transformer (Dosovitskiy et al.,  
1056 2021) and GPT-2 (Radford et al., 2019)). As  
1057 with multi-label classification, we train the mod-  
1058 els with early stopping. Specifically, for Clip-  
1059 Cap, we follow the official implementation regard-  
1060 ing the training settings. For BLIP-2 and Trans-  
1061 former, we use the implementation in Hugging  
1062 Face (Wolf et al., 2020). We use the AdamW opti-  
1063 mizer (Loshchilov and Hutter, 2019) with a learn-  
1064 ing rate of  $2 \times 10^{-6}/1 \times 10^{-4}$  and batch size of  
1065 8/64 for BLIP-2 and Transformer, respectively.

1066 **Results for skin tone.** We show the results of the  
1067 experiments for skin tone bias mitigation in Table 7.  
1068 The results show that the insights in the main paper  
1069 are mostly consistent across the protected groups.

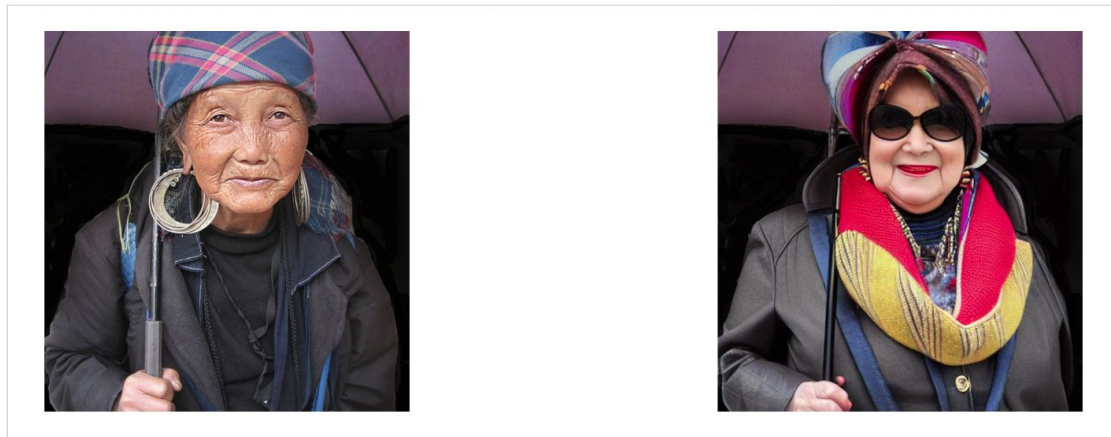
## 1070 B.3 Human Filter Evaluation

1071 In Figures 7 to 9, we present example tasks for  
1072 human evaluation conducted on Amazon Mechan-  
1073 ical Turk (AMT) (Turk, 2012). This evaluation  
1074 assesses how well each combination of filters iden-  
1075 tifies desirable inpainted images. Figure 7 shows  
1076 the user interface for evaluating the similarity of  
1077 held/nearby objects and their colors between the  
1078 original (left) and inpainted (right) images. Fig-  
1079 ure 8 asks workers to select a skin tone class us-  
1080 ing the Monk Skin Tone Scale (Schumann et al.,  
1081 2023; Monk, 2023). We conduct this evaluation on  
1082 both original and inpainted images and compute  
1083 the degree of agreement between them. Figure 9  
1084 verifies if *perceived* gender is accurately depicted—  
1085 according to the AMT worker—in the inpainted  
1086 images through gap-filling, where workers must  
1087 choose a protected group term to complete the sen-  
1088 tence. Each assignment pays \$0.07, with a total  
1089 participant compensation of approximately \$2,000.

	ClipCap			BLIP-2			Transformer		
	M	CS	LIC	M	CS	LIC	M	CS	LIC
Original	<b>29.4</b>	75.3	4.6	<b>27.1</b>	<b>73.9</b>	2.2	<b>27.0</b>	<b>71.5</b>	5.3
$\mathcal{S}_{\text{synthetic}}$ (Ours)	29.1	<b>75.4</b>	<b>3.7</b>	26.8	73.6	<b>2.0</b>	26.5	71.0	<b>4.7</b>

Table 7: Captioning quality and skin tone bias scores of ClipCap, BLIP-2, and Transformer backbones on COCO. M and CS denote METEOR and CLIPScore. **Bold** represents the best. For an unbiased model, Ratio = 1 and LIC = 0.

Please compare the target person (right) with the reference person (left) and answer the questions below.



Q1. Is the target person similar to the reference person? Focus only on **type** of clothing worn by the persons and the **type** of objects they are holding/touching.

Has significant discrepancies  Has minor discrepancies or identical objects

Q2. Is the target person similar to the reference person? Focus only on the **color** of the clothing worn by the people and the **color** of the objects they are holding/touching.

Has significant discrepancies  Has minor discrepancies or indistinguishable colors

Figure 7: Evaluation of *perceived* object and color similarity between original and inpainted images on AMT.

Please answer the following question about the image below.



Q. What is the skin tone of the person? If you are not sure, then select "Unsure".

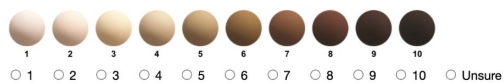


Figure 8: Evaluation of *perceived* skin tone using the Monk Skin Tone Scale on AMT.



Please compare the image and description, and answer the following questions.



Q. Choose the best word to complete the sentence. If you are not sure, then select "Unsure".

- Woman / she / her / hers    Man / he / him / his    Unsure

Submit

Figure 9: Evaluation of *perceived* gender depiction accuracy in inpainted images on AMT.