Auditing Image-based NSFW Classifiers for Content Filtering

Warren Leu*
wwleu@uci.edu
University of California, Irvine
Irvine, USA

Yuta Nakashima
n-yuta@ids.osaka-u.ac.jp
Osaka University
Suita, Japan

Noa Garcia
noagarcia@ids.osaka-u.ac.jp
Osaka University
Suita, Japan

ABSTRACT

This paper examines NSFW (Not Safe For Work) image classifiers for content filtering. Through an audit of three prevalent NSFW classifiers, we analyze the relationship between NSFW predictions and three demographic factors: gender, skin-tone, and age. Our study reveals that women are disproportionately more frequently misclassified as NSFW compared to men, even when they appear conducting common daily-life activities. Additionally, we find that NSFW classifiers tend to mispredict images of people with lighter skin-tones and images depicting younger people. We explore the causes of such mispredictions by analyzing the explanatory pixel maps, which reveal some of the reasons behind the misclassifications. Overall, the implications of our findings become particularly salient when considering the application of filters based on NSFW classifiers, which we identified to have a direct impact on image datasets, computer vision models, generative AI, user experience, and artistic creativity. In summary, we hope our study brings attention to the inherent biases within NSFW classifiers and underscores the importance of addressing these issues to ensure fair and equitable outcomes in content filtering.

CCS CONCEPTS

• **Social and professional topics** → **Pornography**; **Censoring filters**; **Technology and censorship**; *Race and ethnicity*; **Women**; *Men*; *Age*; • **Computing methodologies** → **Computer vision**.

KEYWORDS

audit, computer vision, content filtering, content moderation, NSFW classification

ACM Reference Format:
Warren Leu, Yuta Nakashima, and Noa Garcia. 2024. Auditing Image-based NSFW Classifiers for Content Filtering. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658963

1 INTRODUCTION

Datasets are an integral part in the development and optimization of machine learning products. They serve various purposes, from model training and parameter selection to performance evaluation and benchmarking against other models. Specifically in the field of computer vision, the emergence of large open-source annotated datasets like ImageNet [15], MSCOCO [37], and OpenImages [35], facilitated the advancement of deep learning models that heavily rely on extensive data [25, 34, 44]. In recent years, there has been a surge in the collection of multimodal datasets comprising image and text pairs to meet the escalating demand for vast amounts of data. Whereas some data collections have been made publicly available for anyone to use and scrutiny, such as Google Conceptual Captions [11, 54], RedCaps [16], or LAION [51, 52], others remain confidential and obfuscated. Examples of the latter include ALIGN [30], ALT200M [28], or datasets used for training large multimodal models such as CLIP [44], DALL-E [45, 46], Parti [69], or Imagen [50]. In any way, the collection method of large multimodal datasets consists of automated web crawling, which enables the aggregation of billions of samples. Nevertheless, as the scale increases, the number of challenges related to data grows, including issues about representation, consent, or the presence of toxic content [7–9].

With respect to toxic, offensive, or abusive content, to prevent undesirable samples from becoming part of a dataset, a common approach is using filters during the data collection process. Common filters applied to images include restrictions on their format (e.g., only *jpg* or *png*), size (e.g., more than 5 kilobytes), aspect ratio (e.g., maximum ratio of larger to smaller dimension of 2.5), license (e.g., only *Creative Commons*), provenance (e.g., only images hosted on Flickr[1]), or content filters (e.g., images not flagged by a NSFW classifier). In this work, we are interested in auditing the use of NSFW classifiers for content filtering and analyzing their implications on the datasets and models.

NSFW[2] (Not Safe for Work) is an Internet acronym used to flag content as inappropriate, usually due to its sexual, violent, or otherwise offensive nature. An NSFW classifier refers to a machine learning model specifically designed to identify if a sample, whether an image, a video, or a piece of text, falls into the NSFW category. By using NSFW classifiers for content filtering, sexual, violent, or otherwise offensive content can be ideally identified and removed. While NSFW or toxicity detectors have been extensively studied in the natural language domain for text inputs [3, 13, 20, 42, 43], the efficacy of image-only NSFW classifiers and their limitations are not well-studied. The opaqueness in their training process, stemming from the nature of NSFW images, makes benchmarking these types of classifiers challenging. As far as we know, no study yet delves into the deep aspects of image-based NSFW classifiers, particularly focusing on the correlations between the prediction of NSFW content and demographic factors such as gender, race, or age. We argue that the ramifications of such correlations can inadvertently

*Work conducted during an internship at Osaka University.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658963

[1] https://www.flickr.com
[2] https://en.wikipedia.org/wiki/Not_safe_for_work

perpetuate and amplify prejudices within the filtered content, raising questions about the ethical implications of automated content filtering.

In this paper, we conduct a comprehensive examination of three image-based NSFW classifiers used recently in multimodal datasets and computer vision models for filtering content [16, 49, 52]. The three classifiers use different architectures, with one using standard convolutional neural networks (CNNs), and the other two relying on multimodal CLIP embeddings [44]. As training is conducted by different individuals and institutions, we assume that the three classifiers are trained on different datasets, although not many specific training details are available. We analyze the False Positive Rates (FPR) of each classifier on two evaluation datasets [21, 71] that contain people but are free of NSFW images, i.e., all the images are Safe for Work (SFW), and compare them against a controlled dataset without images of humans [2]. Then, we investigate differences in the FPR across perceived gender, skin-tone, and age. Our findings reveal a concerning trend: women are disproportionately misclassified as NSFW images at a higher probability compared to men. This discrepancy not only underscores the limitations of existing NSFW classifiers but also can amplify the already pronounced representational gap between men and women in digital content [17]. We also found discrepancies in the FPR according to skin-tone, with images of lighter skin-tone people exhibiting higher FPR, and age, with images of younger people exhibiting higher FPR. An in-depth analysis with explainable artificial intelligence (AI) techniques unveils that some of the pixels predominantly contributing to the misclassification of images as NSFW in *all the three NSFW classifiers* are those associated with female faces.

We conclude the paper by examining the repercussions of gender bias in NSFW classifiers for content filtering, raising discussions about its effects on image datasets, computer vision models, generative AI, user experience, and artistic creativity. We aspire that our work raises awareness and promotes discussions regarding the limitations of content filtering algorithms. By shedding light on the complex interplay between visual cues, biases, and the challenges associated with effectively mitigating explicit content in multimodal datasets, we aim to stimulate further exploration in this domain.

2 RELATED WORK

2.1 Toxicity in Image Datasets

In computer vision and machine learning research, the choice of training datasets plays a crucial role in shaping the performance and ethical considerations of models. Unfortunately, several widely used image datasets have been identified to contain toxic and problematic content, ranging from offensive imagery to explicit and non-consensual material. Take, for instance, ImageNet [15], the dataset in image classification that facilitated the emergence and popularity of convolutional neural networks [34]. Despite efforts to curate its labels by removing 1,593 out of 2,832 inappropriate categories from the WordNet [19] person sub-tree [68], subsequent scrutiny by Birhane and Prabhu [8] revealed the persistence of non-consensual and explicit content. In the same work, Birhane and Prabhu [8] uncovered that the Tiny Images dataset [62], containing 80 million low-resolution images sourced from Internet search engines for image classification tasks, included derogatory terms as labels and offensive visual content, leading to its official withdrawal [61]. Similar trends have been observed in large multimodal datasets, such as LAION-400M [52], a 400 million text-image pairs dataset derived from web page alt-text and used to train generative AI models such as Stable Diffusion [49]. LAION-400M, analyzed by Birhane et al. [9], retained problematic images and text pairs depicting rape, racism, and explicit content. Moreover, a recent study [7] shows that dataset scale exacerbates hateful content. In this way, LAION-5B [51], the latest and largest iteration of the LAION datasets with 5 billion text-image pairs, has been recently removed due to the identification of thousands of instances of suspected child sexual abuse material [60]. Other popular multimodal datasets such as the widely-used MSCOCO [37] and Google Conceptual Captions [54], both envisioned for training image captioning models, have been flagged for unbalanced representations in terms of gender and skin-tone [21, 71]. Overall, scrutinizing these datasets highlights the challenges in ensuring ethical, safe, and non-toxic samples in training datasets.

2.2 NSFW Classifiers for Content Filtering

Manually removing toxicity from large image datasets requires a significant amount of resources. Additionally, visually inspecting millions of images to check whether the depicted content is potentially harmful has been found to have detrimental effects on the mental health of annotators [14, 57, 58]. As a result, some authors choose to formally withdraw datasets upon discovering inappropriate content [51, 62]. An alternative approach involves implementing NSFW classifiers to detect and remove explicit or inappropriate content automatically. NSFW classifiers can take the form of various architectures, from CNNs [25, 34, 55] to multimodal approaches that combine text and image information [44]. Moreover, the rise of image generation models [45, 46] has amplified the risk of producing toxic images. In response, some image generation models [49] now incorporate NSFW classifiers to filter outputs that may be considered toxic or inappropriate.

3 METHODOLOGY

Our audit on image-based NSFW classifiers consists of evaluating three different models on two evaluation datasets and a control dataset. Specifically, we evaluate their performance according to the perceived gender, skin-tone, and age of the people in the images and study disparities in their misclassification rates produced by such demographic factors.

3.1 NSFW Classifiers under Evaluation

We analyze three NSFW classifiers that have recently been used for filtering inappropriate content from either datasets or AI-generated images. We specifically select these three NSFW classifiers due to their presence in state-of-the-art computer vision research, being an indication that they have transcended theoretical frameworks and are actively integrated into practical applications.

The selected NSFW classifiers, namely *NSFW-CNN*, *CLIP-classifier*, and, *CLIP-distance*, and their main characteristics are summarized in Table 1. *NSFW-CNN* extracts embeddings from images with a CNN [59], whereas *CLIP-classifier* and *CLIP-distance*

Table 1: NSFW classifiers in our audit. Underline indicates the parts of the models that need training, *Data source* from where the training data was collected, and *Num. samples* the number of samples used for such training.

	Model	Outputs	Size (MB)	Trained	Data source	Num. samples	Used in
NSFW-CNN	InceptionV3	5	85.30	Yes	NSFW data scraper [33]	Unknown	[8, 16]
CLIP-classifier	CLIP + FC classifier	1	888.32	Yes	LAION-5B subset [51]	$\sim 242,000$	[51]
CLIP-distance	CLIP + cosine distance	17	887.52	No	-	-	[49, 52]

use pre-trained CLIP embeddings [44]. From the CLIP embeddings, *CLIP-classifier* predicts NSFW content with a three-layer fully connected (FC) network trained on LAION-5B dataset [51], while *CLIP-distance* does not require training and classifies images as NSFW if the resulting cosine similarity between their embeddings and a set of pre-defined concepts is above a threshold. The specific technical details for each method are provided below:

NSFW-CNN An InceptionV3 [59] CNN model from [36] trained to classify images into five categories: *sexy*, *neutral*, *porn*, *hentai*, and *drawings*. Given an input image, the *NSFW-CNN* outputs a score from 0 to 1 for each category. If the score corresponding to *porn* is higher than 0.7, the image is classified as NSFW. The model is trained end-to-end with data collected from an NSFW data scraper [33], although the amount of training samples is not specified. This model has been used in Birhane and Prabhu [8] for abusive content detection and in the RedCaps dataset [16] for content filtering.

CLIP-classifier A CLIP image encoder [44] followed by a three-layer FC classifier. The CLIP encoder is a frozen ViT L/14 network [12], and the FC classifier is trained on a subset[3] of the LAION-5B dataset [51]. The *CLIP-classifier* outputs a single 0 to 1 score representing the confidence of the image being NSFW, with 1 being NSFW. If the score is higher than 0.7, the image is classified as NSFW. The LAION organization[4] supplied the model alongside the LAION-5B dataset. However, the model was not used for filtering content in LAION-5B. Instead, it was offered to assist users in filtering the data at their discretion.

CLIP-distance A CLIP image encoder [44] followed by a distance computation. The CLIP encoder is also a frozen ViT L/14 network [12] that converts an input image to an image embedding. The distance between the image embedding and a set of 17 precalculated text embeddings, each representing an NSFW concept, is computed. If the cosine distance between the image embedding and any of the precomputed text embeddings is over a set threshold, the image is classified as NSFW. Note that the details of the specific 17 concepts have not been revealed. This model, which does not require training, can be found inside Stable Diffusion v1.5 [49] by HuggingFace [63] as a safety checker, to detect if a generated image is NSFW, and if so, returning a blacked out image instead of the generated one. A similar approach is used in LAION-400M dataset [52] for content filtering.

[3]Details on how the images in the subset were chosen are not specified by their authors.
[4]https://laion.ai

3.2 Evaluation Datasets

We evaluate the three NSFW classifiers on two annotated subsets of popular image datasets: the GCC dataset [54] with PHASE annotations [21] and the MSCOCO dataset [37] with Zhao et al.'s annotations [71]. The details of each dataset are provided below:

GCC The Google Conceptual Captions (GCC) dataset [54] is a collection with about 3 million text-image pairs automatically collected from the Internet and split into training and validation sets. From those, 18,889 images are manually annotated in PHASE [21] by labeling people in the images according to six perceived attributes: *age*, *gender*, *skin-tone*, *ethnicity*, *emotion*, and *activity*, with a total of 35,347 annotated people and a highly unbalanced class distribution. We use annotations on binary perceived gender (*woman*, *man*), binary perceived skin-tone (*lighter* skin-tone, *darker* skin-tone), and four perceived age categories (*child* (0-14 years old), *young* (15-29 years old), *adult* (30-64 years old), and *senior* (65 years old or more)). We only use images in which all the people have the same perceived attributes, e.g., all woman. Statistics about the number of samples for each class are provided in Table 2.

MSCOCO The Microsoft Common Objects in Context (MSCOCO) dataset [37] is a collection with about 200,000 images labeled with objects, keypoints, and captions. From those, 15,762 images are manually annotated with perceived *gender* and *skin-tone* attributes by Zhao et al. [71], with a total of 28,315 annotated people. Similar to GCC, we use annotations on binary perceived gender (*woman*, *man*), and binary perceived skin-tone (*lighter* skin-tone, *darker* skin-tone). Perceived age is not available in this dataset. Images annotated with "both" and "unsure" are ignored. Statistics about the number of samples per class are provided in Table 2.

3.3 Control Dataset

Additionally, we use a control dataset with images without people: the PASS dataset [2].

PASS The Pictures without humAns for Self-Supervision (PASS) dataset [2] is a collection with about 1.4 million unlabeled images that do not include any pictures of humans. It is designed to prevent issues with privacy, data protection, and ethics. We use PASS as a control benchmark to examine how each of the NSFW models performs when given images that do not contain people. To compare results fairly in terms of scale, we use a random subset of 11,685 images.

Table 2: Number of images used in our analysis per dataset and attribute.

	Gender			Skin-tone			Age				
	Woman	Man	Total	Light	Dark	Total	Child	Young	Adult	Senior	Total
GCC	4,037	7,069	11,106	8,607	1,048	9,655	745	2,535	2,980	223	6,483
MSCOCO	3,611	8,017	11,628	10,635	1,358	11,993	-	-	-	-	-

3.4 Experimental Details

We run our experiments on Python 3.11.5 with PyTorch 2.1 [40] and TensorFlow 2.14.0 [1] on a single GeForce RTX 3070 GPU. We do not re-train any of the three NSFW classifiers, but use them off-the-shelf as provided by their authors. Input images are resized to 299×299 pixels for *NSFW-CNN* and 224×224 pixels for both *CLIP-classifier* and *CLIP-distance*.

4 IMAGE-BASED NSFW CLASSIFIERS AUDIT

We conduct our audit in four phases. In the initial phase (Section 4.1), we benchmark the three NSFW classifiers by comparing their performance across the evaluation and control datasets. In the second phase (Sections 4.2, 4.3, and 4.4), our focus shifts towards analyzing the relationship between demographic attributes and NSFW predictions: Section 4.2 focuses on gender, Section 4.3 on skin-tone, and Section 4.4 on age. Moving forward to the third phase (Section 4.5), we investigate the specific image regions triggering the NSFW classifiers by exploring pixel importance maps generated with explainable AI tools. In our final analysis (Section 4.6), we discuss the implications of the relationship between demographics and NSFW misclassification rates.

4.1 Comparative Evaluation

Firstly, we compare the performance of the three NSFW classifiers on the two evaluation datasets and the control dataset. Specifically, the performance of each NSFW classifier is measured as the False Positive Rate (FPR). Given an image I as input, an NSFW classifier, C, which gives a confidence value, or how likely I is being NSFW, predicts whether the image is NSFW or not as

$$\hat{y}(I) = \begin{cases} 1 & \text{if } C(I) > th \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

with $\hat{y} = 1$ if the image is predicted as NSFW, and $\hat{y} = 0$ otherwise, where th is a predefined threshold. As all the samples in the evaluation and the control datasets are safe for work (SFW), their ground truth label, y, is always 0. The FPR is computed as the number of incorrectly predicted NSFW images over dataset $\mathcal{D}$ (either control, GCC, or MSCOCO) as

$$\text{FPR} = \frac{1}{|\mathcal{D}|} \sum_{I \in \mathcal{D}} \hat{y}(I), \tag{2}$$

where $|\mathcal{D}|$ gives the number of images in $\mathcal{D}$.

Note that a low FPR is not always desirable, especially if achieved at the expense of a high False Negative Rate (FNR), which may lead to the classification of numerous inappropriate images as safe. Nevertheless, FNR is not computed due to a lack of properly annotated NSFW datasets. Thus, our analysis centers on comparing the performance of the NSFW classifiers on the evaluation datasets with their behavior on the control set. For completeness, we also report the average confidence score on the NSFW classifiers, given as

$$\frac{1}{|\mathcal{D}|} \sum_{I \in \mathcal{D}} C(I). \tag{3}$$

Results are shown in Table 3. On the control dataset, *CLIP-classifier* achieves the lowest FPR, with only a single image misclassified as NSFW, while *NSFW-CNN* and *CLIP-distance* misclassify 10 and 131 images, respectively. The FPRs for *NSFW-CNN* and *CLIP-distance* on MSCOCO are similar to the control set, but they substantially increase on the GCC dataset. The GCC evaluation set seems to contain images that are generally more challenging for all the models to classify. Among the three datasets, GCC has the more lenient collection method, potentially resulting in a higher frequency of NSFW-like images. Finally, the *CLIP-classifier* performance is substantially different between the two evaluation sets featuring people, MSCOCO and GCC, and the control set without people. The FPR increasing from 0.009 in the Control dataset up to 7.509 in the GCC dataset indicates a strong correlation between images of people and NSFW content within the internal representations of this model.

4.2 Gender Examination

Next, we examine how the perceived gender in input images influences the predictions made by the NSFW classifiers. As both GCC and MSCOCO datasets are unbalanced and contain more images from man than woman, we compute the FPR per gender, FPR_g with gender $g \in \{woman, man\}$, as

$$\text{FPR}_g = \frac{1}{|\mathcal{D}_g|} \sum_{I \in \mathcal{D}_g} \hat{y}(I). \tag{4}$$

where $\mathcal{D}_g \subset \mathcal{D}$ only contains images with gender g.

Results are shown in Table 4. Images with perceived women are misclassified as NSFW at higher rates than images with perceived men. The difference is disproportionately high for the case of *CLIP-classifier* on the GCC dataset, reaching an alarming margin of 17.9%. The gender disparity, which appears in the three NSFW classifiers, is more pronounced in the GCC dataset than in the MSCOCO dataset. Some examples are shown in Figures 1 and 2 for MSCOCO and GCC datasets, respectively. Upon inspecting the images, we find that most pictures of women depict them engaging in innocuous and common activities like sports, eating, or posing for a camera. For men, a significant portion of the limited number of images classified as NSFW showcases characteristics associated with femininity or gender nonconformity. This suggests that NSFW classifiers tend to

Table 3: NSFW classifiers comparative evaluation. NSFW indicates the number of images misclassified as NSFW, FPR the false positive rate in %, and score the average confidence score for each classifier. Bold font highlights the classifier with the highest mispredictions and FPR for each dataset.

	Control (11, 685 images)			GCC (11, 106 images)			MSCOCO (11, 628 images)		
	NSFW	FPR (%)	score	NSFW	FPR (%)	score	NSFW	FPR (%)	score
NSFW-CNN	10	0.397	0.086	68	0.612	0.032	49	0.396	0.021
CLIP-classifier	1	0.009	0.001	**834**	**7.509**	0.106	**249**	**2.141**	0.045
CLIP-distance	**131**	**1.121**	-	520	4.682	-	144	1.238	-

Table 4: False Positive Rate per gender in percentage (%). Diff. is the difference between Woman and Man columns. Bold font highlights the gender with the highest mispredictions for each classifier and dataset.

		GCC			MSCOCO		
	Control	Woman	Man	Diff.	Woman	Man	Diff.
NSFW-CNN	0.397	**1.248**	0.211	1.037	**0.832**	0.237	0.594
CLIP-classifier	0.009	**18.530**	0.660	17.871	**5.123**	0.798	4.325
CLIP-distance	1.121	**5.545**	4.112	1.433	**1.246**	1.235	0.011

categorize an image as NSFW based on the presence of traditionally associated feminine traits.

4.3 Skin-Tone Examination

We analyze the relationship between skin-tone and NSFW predictions. We compute the FPR per skin-tone, FPR_s with skin-tones $s \in \{darker, lighter\}$, as

$$\text{FPR}_s = \frac{1}{|\mathcal{D}_s|} \sum_{I \in \mathcal{D}_s} \hat{y}(I). \tag{5}$$

where $\mathcal{D}_s \subset \mathcal{D}$ is the subset of images annotated with skin-tone s.

Results are shown in Table 5. Notably, all three classifiers exhibit a higher rate of false positives for images featuring individuals with perceived lighter skin-tones compared to those with darker skin-tones. In line with the analysis of gender bias, the *CLIP-classifier* on the GCC dataset shows the most substantial difference in the FPR, although the disparities are less pronounced than in the gender evaluation. Note that the number of images per class is more unbalanced than in gender, with about 8 times more individuals of lighter skin-tones than darker skin-tones. Regardless, these results suggest that skin-tone may not be as robust an indicator for NSFW classifiers as gender.

4.4 Age Examination

The last demographic attribute we analyze is age. Similarly to gender and skin-tone, we compute the FPR per age, FPR_a, over $\mathcal{D}_a$ with age $a \in \{child, young, adult, senior\}$, as

$$\text{FPR}_a = \frac{1}{|\mathcal{D}_a|} \sum_{I \in \mathcal{D}_a} \hat{y}(I). \tag{6}$$

where $\mathcal{D}_a \subset \mathcal{D}$ is the subset of images annotated with age a.

Results are shown in Table 6, only for the GCC dataset, as MSCOCO dataset does not contain age annotations. FPR is well above the control dataset for all the age groups and classifiers. Of particular concern is the observation that the age groups most prone to misclassification are those associated with younger individuals. The *Child* category (0-14 years old) exhibits the highest rate of mispredicted NSFW in both the *NSFW-CNN* and *CLIP-classifier*, while the *Young* category (15-29 years old) has the highest mispredicted NSFW rate in the *CLIP-distance* classifier. For all models, many of the child images classified as NSFW are images of babies without clothes or just in diapers. This suggests that exposed skin may play a factor in classification, as will also be seen later in Section 4.5. Why children have a higher rate of false positives, however, is still unclear.

4.5 Regional Analysis

Our next evaluation involves looking into the NSFW classification mechanisms and understanding which particular regions of the image trigger the NSFW prediction.

NSFW-CNN regional analysis. We analyze the contribution of each region to the final prediction through Grad-CAM [53]. Grad-CAM is an explainable AI algorithm that generates heatmaps in the original image, representing the regions that have the most influence on the final prediction - in our case, whether the image is classified as NSFW or not. Some examples, with confidence above 0.9, are shown in Figure 3. We note the following observations:

(1) Images misclassified as NSFW often depict individuals, especially those annotated as women, engaged in eating. The reason for this classification is unclear; it remains uncertain whether the model associates eating or open mouths with sexual content or if it is influenced by the prevalence of close-up shots of faces in these images.
(2) Another category of frequently misclassified NSFW images involves hands, with the specific reason behind this misclassification also remaining unclear. Common to both types of misclassifications is the belief that a significant amount

Figure 1: Examples of images from the MSCOCO dataset misclassified as NSFW per classifier. On the top row (orange background), images annotated as man. On the bottom row (purple background), images annotated as woman. Note that annotations are based on perceived gender.

Figure 2: Examples of images from the GCC dataset misclassified as NSFW per classifier. On the top row (orange background), images annotated as man. On the bottom row (purple background), images annotated as woman. Note that annotations are based on perceived gender.

of exposed skin in the image plays a substantial role in the model's decision-making process.

(3) In the last image in Figure 3, there is a noticeable focus on the face of the person, which is annotated as woman. The emphasis on facial features is a recurring theme evident when examining the heatmaps for the *CLIP-classifier* model.

Table 5: False Positive Rate per skin color in percentage (%). Diff. is the difference between the columns Light (skin-tone) and Dark (skin-tone). Bold font highlights the skin-tone with the highest mispredictions for each classifier and dataset.

		GCC			MSCOCO		
	Control	Light	Dark	Diff.	Light	Dark	Diff.
NSFW-CNN	0.397	**0.732**	0.095	0.637	**0.489**	0.221	0.268
CLIP-classifier	0.009	**8.354**	1.908	6.445	**2.351**	1.178	1.173
CLIP-distance	1.121	**5.193**	1.622	3.571	**1.590**	0.736	0.854

Table 6: False Positive Rate per age in percentage (%). Bold font highlights the age with the highest mispredictions for each classifier and dataset.

		GCC			
	Control	Child	Young	Adult	Senior
NSFW-CNN	0.397	**1.074**	0.907	0.369	0.000
CLIP-classifier	0.009	1.208	**12.308**	3.523	0.000
CLIP-distance	1.121	**9.128**	6.785	3.993	5.830

(a) Eating (b) Hands (c) Faces

Figure 3: *NSFW-CNN* classifier regional analysis conducted with Grad-CAM [53]. We find three main themes within the misclassified images: (a) people eating, (b) hands, and (c) women's faces. Red regions indicate a higher contribution to the model prediction, whereas blue regions indicate a low contribution.

CLIP-classifier regional analysis. In this case, we use RISE [41] to obtain pixel-level explanations of the regions with the highest contribution to the NSFW prediction. RISE is a method for empirically estimating pixel importance by masking random regions of the image and observing the differences in the model prediction. Examples of RISE heatmaps for *CLIP-classifier* are shown in Figure 4. The most notable observations can be summarized as follows:

(1) Images misclassified as NSFW by the *CLIP-classifier* share many traits with the *NSFW-CNN* model, such as images annotated as women being overwhelmingly more likely to be classified as NSFW than men. Another similarity is the tendency to see more exposed skin as NSFW, though it seems to be a smaller factor here.
(2) A common element seen in almost every picture is that the pixel-level explanations are focused on the area of the face. This is present even in images with a more sexually explicit tone, showing that this model seems to mainly use faces in images to classify NSFW or not.
(3) In the last image in Figure 4, the pixel contribution is focused around the woman's facial region, despite the image having a much more exposed man right beside. This seems to suggest that not only does the model tend toward faces when classifying, but it tends specifically toward feminine faces, which is supported by the overwhelming majority of images classified as NSFW being annotated as women.

CLIP-distance regional analysis. For this model, we also use RISE to obtain the explanations for the predictions. In this case, we identify four themes within the misclassified images: faces, sausages, donuts, and eating. Examples are shown in Figure 5 and the most notable observations are summarized as:

(1) We find heatmaps focusing on the facial regions of women. However, this case is not as prominent as in *NSFW-CNN* and *CLIP-classifier*, so we believe that facial features have a present but less pronounced effect in the *CLIP-distance* model.
(2) Another theme that is shared with the previous models is classifying images of people eating as NSFW. Unlike the previous two models, it does not seem that *CLIP-distance* associates eating with an image being NSFW. Instead, it may see the object itself as NSFW. This can be seen from

Figure 4: *CLIP-classifier* regional analysis conducted with RISE [41] by estimating pixel importance in the input image. We find that the model focuses especially on women's faces when mispredicting safe images as NSFW. Red regions indicate a higher contribution to the model prediction, whereas blue regions indicate a low contribution.

comparing the explanation maps of people eating: whereas the *NSFW-CNN* model focuses on the mouth and facial area, the *CLIP-distance* model focuses mainly on the object.

(3) Several NSFW images are of sausages or bananas. Another food item the model predicts as NSFW is donuts, specifically the donut hole area. Considering that *CLIP-distance* relies on embedding distances between images and textual embeddings from unknown NSFW concepts, it looks like some of those concepts may represent shapes similar to those items. In this sense, the model is not capable of discerning between common food objects and NSFW content. It is important to note that, for all of the images with only food, neither *NSFW-CNN* nor *CLIP-classifier* classify them as NSFW, being this behavior specific to *CLIP-distance* only.

4.6 Implications for Content Filtering

Finally, we analyze the implications of the above results, particularly when NSFW classifiers are used for filtering content in datasets, generative AI images, or social media platforms. Our examination revolves around five issues: the impact on image datasets, the impact on computer vision models, the impact on generative AI, the implications for user experience, and the implications for artistic creativity.

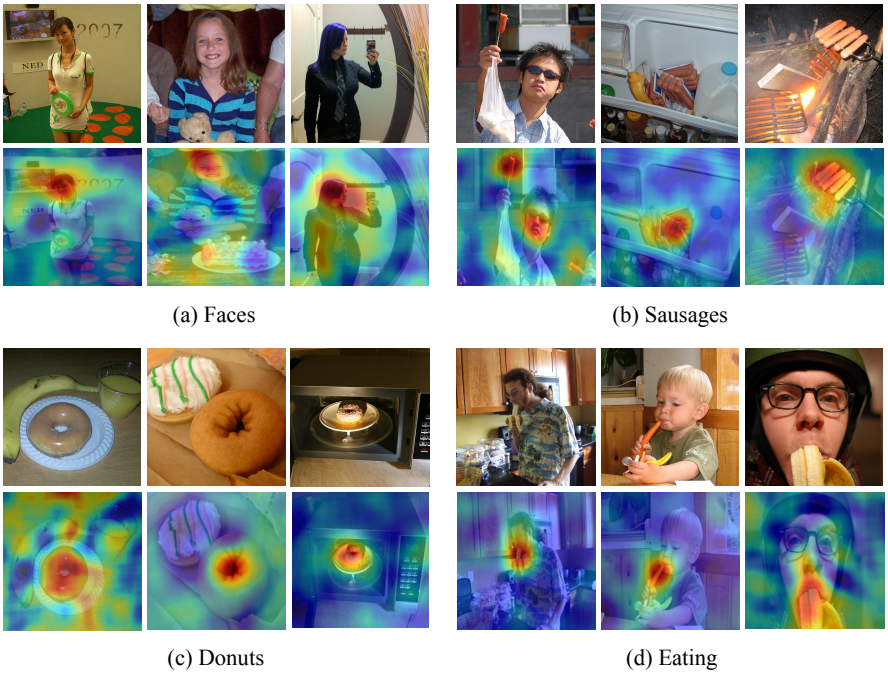
Impact on Image Datasets. A persistent issue in terms of social bias in image datasets is the disparities in the representational gap for different demographic groups [21, 71]. For example, for gender bias, the quantity of images depicting women tends to be significantly smaller compared to those of men. In addition to the already analyzed MSCOCO and GCC datasets, which exhibit 2.22 and 1.64 times more men than women, respectively, according to [26] the ratio of men to women in visual question answering (VQA) datasets ranges from 1.7 in GQA [29] to 2.1 in VQA 2.0 [23] and Visual7W [73]. Using an NSFW classifier to filter content during the dataset creation phase, coupled with the higher likelihood of images featuring women being misclassified as NSFW, can exacerbate the representational gap and increase the already high ratio of men to women in computer vision datasets.

Impact on Computer Vision Models. Our findings hold a direct impact on the performance of computer vision models, which undergo training on large multimodal datasets filtered through NSFW classifiers [44–46, 50, 69]. Models trained on unbalanced datasets not only mirror the biases present in the original data but also have the potential to amplify them [24, 27, 64, 66, 70, 72], leading to the generation of skewed predictions at elevated rates. Recent research [65] underscores the importance of data in mitigating the adverse effects of bias, highlighting the need for careful considerations in training data selection to foster fair and accurate model outcomes.

Impact on Generative AI. With ongoing discussions about the ethical implications of image generation models [4, 6, 31, 32], including bias [5, 38, 67], intellectual property [56], and privacy [10], some efforts to address the generation of toxic or inappropriate images have involved the inclusion of NSFW classifiers for posthoc deletion of generated images [49]. However, as our study reveals, these classifiers exhibit a higher rate of misclassification for images containing women compared to those featuring men, which implies that the use of NSFW classifiers may reduce the final production of images depicting women, exacerbating the existing representational gap within these models. As the field advances, it becomes necessary to examine the consequences of such measures and advocate for a more equitable and inclusive trajectory in generative AI development.

Implications for User Experience. The use of automatic tools for content moderation in social media platforms has been largely discussed [22], especially for text data. When applied to visual content, NSFW models could disproportionately remove images and videos featuring women compared to men. As shown in Section 4.5, female faces undergo higher NSFW misclassification rates, which may lead social media users to encounter fewer images of women in comparison to their male counterparts. This skewed visibility may inadvertently foster a misleading impression that women are less prevalent in society. The unintended consequence of such content filtering mechanisms could contribute to distorted perceptions of gender representation within the online environment.

Implications for Artistic Creativity. Algorithmic content moderation on social media has a direct influence on the creativity of artists, directly impacting the visibility of their work and their income [39, 47, 48]. The results presented in this paper, where NSFW algorithms flag content based on the presence of female faces, add further evidence to the growing concerns about automatic content moderation algorithms censoring artistic pieces featuring the female body [18], even when the intent is purely artistic and non-sexual.



(a) Faces (b) Sausages

(c) Donuts (d) Eating

Figure 5: *CLIP-distance* regional analysis conducted with RISE [41]. In this case, we find four groups of images that often flag the NSFW misprediction: (a) women's faces, (b) sausages, (c) donuts, and (d) people eating. Red regions indicate a higher contribution to the model prediction, whereas blue regions indicate a low contribution.

5 CONCLUSION

In this work, we analyzed three prominent not safe for work (NSFW) classifiers and their impact when used for content filtering and automatic content moderation. We conducted an analysis on the GCC and MSCOCO datasets with demographic annotations and compared the false positive rate (FPR) against a control dataset without humans (PASS dataset). By inspecting the regions with the highest contribution to the NSFW mispredictions, we concluded:

- NSFW classifiers mispredicted images of perceived women at higher rates than images of perceived men. The difference was as high as 17.9%. Upon inspection of the mispredicted images, women appeared doing standard activities like sports, eating, or posing for the camera. For men, we identified a number of images exhibiting gender-nonconforming attributes, indicating that the mere presence of attributes perceived as feminine or non-masculine can flag NSFW classifiers.
- NSFW classifiers mispredicted images of people with lighter skin-tone at higher rates than images of people with darker skin-tone. These results, however, should be considered cautiously due to the big unbalance in the number of samples, with 8 times more light skin-tone annotations than dark skin-tone annotations.
- NSFW classifiers tended to mispredict younger people at higher rates than older people. We found especially concerning the result for the *child* category (0-14 years old), with two out of three classifiers exhibiting the highest FPR and the third one exhibiting the second highest FPR. This indicates that NSFW classifiers find NSFW traits in innocuous images of children, inducing reflection about the training data and the reasons why images showing children and younger adults activated NSFW predictions.
- When conducting a regional analysis, we found that all three NSFW classifiers tended to focus on the faces of women to make their prediction. We found that faces of women were a stronger signal for NSFW classifiers than images of men's nude torso.
- The regional analysis also showed that hands and people eating were more likely to raise NSFW flags. Additionally, the NSFW classifier based on distance embeddings (*CLIP-distance*) had more difficulties distinguishing between safe and NSFW objects.
- Finally, by analyzing the impact of demographic biases on NSFW classifiers, we found that different FPR across different demographic groups has the potential to widen the representational gap in image datasets, computer vision models, generative AI, and online content in general. This has a direct impact on users' experience as well as on artistic creativity.

6 LIMITATIONS

Our results about NSFW classifiers and their implications for content filtering, while insightful, have certain limitations that warrant consideration. The exclusive use of a single metric, namely False

Positive Rate (FPR), may present an incomplete picture of the classifier's overall performance. A comprehensive evaluation should ideally incorporate multiple metrics to ensure a nuanced understanding of its effectiveness. Nevertheless, the decision to refrain from computing accuracy and False Negative Rate (FNR) on NSFW datasets was deliberated, driven by both a lack of reliable NSFW datasets and ethical considerations related to the download and possession of NSFW data. Another limitation of this work is the use of imbalanced demographic annotations, a factor that can introduce noise and skew the results. The unbalanced nature of demographic data in computer vision datasets underscores the need for a more balanced and representative dataset to draw robust conclusions.

ACKNOWLEDGMENTS

This work is partly supported by JST CREST Grant No. JPMJCR20D3, JST FOREST Grant No. JPMJFR216O, JSPS KAKENHI Nos. JP22K12091 and JP23H00497. The JST and JSPS had no role in the design and conduct of the study; access and collection of data; analysis and interpretation of data; preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. The authors declare no other financial interests.

RESEARCHERS POSITIONALITY

The authors acknowledge the importance of using automatic filters to eliminate toxic content from datasets and online platforms. Our intention is not to discourage the use of such filters, which we deem necessary. Rather, with this work, our objective is to highlight the disparities in the functionality of NSFW classifiers across diverse demographic groups. By bringing attention to this issue, we aim to contribute to the collective efforts to address and rectify these disparities, ensuring a more equitable and effective application of content filtering mechanisms.

ETHICAL CONSIDERATIONS

An inherent ethical consideration of this work lies in defining what constitutes an NSFW image. The ambiguity surrounding the NSFW criteria prompts a critical examination of the ethical dimensions involved. Questions arise regarding the threshold for explicit content — what specific body parts, if exposed, classify an image as NSFW? Moreover, the consideration of cultural and contextual variations adds another layer of complexity. The ethical discourse extends to instances where certain body parts may be depicted in art, statues, or classic works, challenging the universality of NSFW categorization. Within the scope of the paper, we find it important to acknowledge these ethical considerations and the diversity of perspectives to ensure a balanced and culturally sensitive approach.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Yuki M. Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. 2021. PASS: An ImageNet replacement for self-supervised pretraining without humans. *NeurIPS Track on Datasets and Benchmarks* (2021).

[3] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proc. 13th International Workshop on Semantic Evaluation*. 54–63.

[4] Oliver Bendel. 2023. Image synthesis from an ethical perspective. *AI & SOCIETY* (2023), 1–10.

[5] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1493–1504.

[6] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Conference on AI, Ethics, and Society*. 396–410.

[7] Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. 2023. Into the LAIONs Den: Investigating Hate in Multimodal Datasets. In *Proc. NeurIPS 2023 Datasets and Benchmarks track*.

[8] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *Proc. WACV*. IEEE, 1536–1546.

[9] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[10] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*. 5253–5270.

[11] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proc. CVPR*. 3558–3568.

[12] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2818–2829.

[13] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. CONAN-COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. In *Proc. ACL*. 2819–2829.

[14] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. In *Proc. AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 33–42.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*. Ieee, 248–255.

[16] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. RedCaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431* (2021).

[17] Masoomali Fatehkia, Ridhi Kashyap, and Ingmar Weber. 2018. Using Facebook ad data to track the global digital gender gap. *World Development* 107 (2018), 189–209.

[18] Gretchen Faust. 2017. *Hair, blood and the nipple*. Vol. 34. transcript Verlag.

[19] Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

[20] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30.

[21] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. 2023. Uncurated image-text datasets: Shedding light on demographic bias. In *Proc. CVPR*. 6957–6966.

[22] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020), 2053951719897945.

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proc. CVPR*. 6904–6913.

[24] Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. 2022. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706* (2022).

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

[26] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Gender and racial bias in visual question answering datasets. In *Proc. FAccT*. 1280–1292.

[27] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 13450–13459.

[28] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proc. CVPR*. 17980–17989.

[29] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proc. CVPR*. 6700–6709.

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. ICML*. PMLR, 4904–4916.

[31] Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. AI Art and its Impact on Artists. In *Conference on AI, Ethics, and Society*. 363–374.

[32] Amelia Katirai, Noa Garcia, Kazuki Ide, Yuta Nakashima, and Atsuo Kishimoto. 2023. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. *arXiv preprint arXiv:2311.18345* (2023).

[33] Alex Kim. [n. d.]. NSFW Data Scraper. https://github.com/alex000kim

[34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Proc. NeurIPS* 25 (2012).

[35] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *Trans. IJCV* (2020).

[36] Gant Laborde. [n. d.]. Deep NN for NSFW Detection. https://github.com/GantMan/nsfw_model

[37] 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 *Proc. ECCV*.

[38] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

[39] Henry Lydiate. 2021. Censorship: Don't Delete Art. *Art Monthly* 445 (2021), 46–46.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Proc. NeurIPS* 32 (2019).

[41] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC)*. http://bmvc2018.org/contents/papers/1064.pdf

[42] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* 55 (2021), 477–523.

[43] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *Proc. EMNLP-IJCNLP*. 4755–4764.

[44] 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 *International Conference on Machine Learning*. PMLR, 8748–8763.

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125* (2022).

[46] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proc. ICML*. PMLR, 8821–8831.

[47] Piera Riccio, Thomas Hoffmann, and Nuria Oliver. 2024. Exposed or Erased: Algorithmic Censorship of Nudity in Art. In *CHI*. 359–363.

[48] Piera Riccio, Jose Luis Oliver, Francisco Escolano, and Nuria Oliver. 2022. Algorithmic Censorship of Art: A Proposed Research Agenda.. In *ICCC*. 359–363.

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*. 10684–10695.

[50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS* 35 (2022), 36479–36494.

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Proc. NeurIPS* 35 (2022), 25278–25294.

[52] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshop Datacentric AI*.

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

[54] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. ACL*.

[55] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[56] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6048–6058.

[57] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17, 4 (2023).

[58] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proc. CHI*. 1–14.

[59] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[60] David Thiel. 2023. Identifying and Eliminating CSAM in Generative ML Training Data and Models. (2023).

[61] Antonio Torralba, Rob Fergus, and Bill Freeman. 2020. *Tiny Images dataset*. https://groups.csail.mit.edu/vision/TinyImages/ Accessed: 2024-01-15.

[62] Antonio Torralba, Rob Fergus, and William T Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Trans. PAMI* 30, 11 (2008), 1958–1970.

[63] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. [n. d.]. *Diffusers: State-of-the-art diffusion models*. https://github.com/huggingface/diffusers

[64] Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In *International Conference on Machine Learning (ICML)*. PMLR, 10882–10893.

[65] Angelina Wang and Olga Russakovsky. 2023. Overwriting Pretrained Bias with Finetuning Data. In *International Conference on Computer Vision (ICCV)*.

[66] 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 *International Conference on Computer Vision (ICCV)*. 5310–5319.

[67] Yankun Wu, Yuta Nakashima, and Noa Garcia. 2023. Stable Diffusion Exposed: Gender Bias from Prompt to Image. *arXiv preprint arXiv:2312.03027* (2023).

[68] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proc. FAccT*. 547–558.

[69] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *TMLR* (2022).

[70] Dora Zhao, Jerone Andrews, and Alice Xiang. 2023. Men also do laundry: Multi-attribute bias amplification. In *Proc. ICML*. PMLR, 42000–42017.

[71] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and Evaluating Racial Biases in Image Captioning. In *Proc. ICCV*.

[72] 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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2979–2989.

[73] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proc. CVPR*. 4995–5004.