Uncurated Image-Text Datasets: Shedding Light on Demographic Bias

Noa Garcia Yusuke Hirota Yankun Wu Yuta Nakashima
{noagarcia@, y-hirota@is., yankun@is., n-yuta@}ids.osaka-u.ac.jp
Osaka University

Abstract

The increasing tendency to collect large and uncurated datasets to train vision-and-language models has raised concerns about fair representations. It is known that even small but manually annotated datasets, such as MSCOCO, are affected by societal bias. This problem, far from being solved, may be getting worse with data crawled from the Internet without much control. In addition, the lack of tools to analyze societal bias in big collections of images makes addressing the problem extremely challenging.

Our first contribution is to annotate part of the Google Conceptual Captions dataset, widely used for training vision-and-language models, with four demographic and two contextual attributes. Our second contribution is to conduct a comprehensive analysis of the annotations, focusing on how different demographic groups are represented. Our last contribution lies in evaluating three prevailing vision-and-language tasks: image captioning, text-image CLIP embeddings, and text-to-image generation, showing that societal bias is a persistent problem in all of them.

https://github.com/noagarcia/phase

1. Introduction

The training paradigm in vision-and-language models has shifted from manually annotated collections, such as MS-COCO [30] and Visual Genome [27], to massive datasets with little-to-none curation automatically crawled from the Internet [17, 42, 43]. Figure 1 illustrates this tendency by comparing the size of paired image-text datasets over time. Whereas manually annotated datasets, widely used in the last decade, contained a few hundred thousand images each, the latest automatically crawled collections are composed of several million samples. This large amount of data has led to training some disruptive models in the field, such as CLIP [37] trained on 400 million image-text pairs; Imagen [41] trained on 860 million image-text pairs; Flamingo [1] trained on 2.3 billion images and short videos paired with text; DALL-E 2 [38] trained on 650 million images; or Stable Diffusion [39], trained on 600 million captioned images. Those models have been shown to learn visual and language representations that outperform the previous state-of-the-art on tasks such as zero-shot classification [37] or text-to-image generation [38, 39].

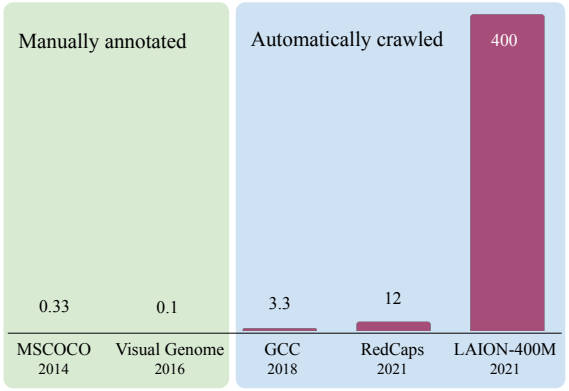

Figure 1. Evolution of paired image-text datasets in terms of number of samples (in million). Datasets scaled up with data automatically crawled from the Internet, reaching the current status in which models are trained with hundreds of millions of samples.

Despite the impressive results on controlled benchmarks, a critical drawback arises: the larger the training set, the less control over the data. With toxic content easily accessible on the Internet, models trained under uncurated collections are more prone to learn harmful representations of the world, including societal bias, which results in models performing differently for different sociodemographic groups [57]. The risk of obtaining unfair representations is high, as not only do models trained on biased datasets learn to reproduce bias but also amplify it by making predictions more biased than the original data [22, 53, 56]. This turns out to be harmful when, far from controlled research environments, models are used in the real-world [10].

Manually annotated datasets [16, 30] have been shown to be affected by societal bias [21, 32, 60, 61], but the problem gets worse in automatically crawled datasets [8, 9]. To overcome societal bias, fairness protocols must be included both in the dataset and in the model development phase. Data analysis [8, 9, 21, 32, 52, 58], evaluation metrics [22, 40, 53],

Table 1. Image-text datasets with annotations for bias detection.

Dataset	Image source	Annotation process	Labels	Images	Regions	Attributes
Zhao et al. [61]	MSCOCO [30]	automatic (captions)	image	33,889	-	gender
Zhao et al. [60]	MSCOCO [30] (val)	crowd-sourcing	region	15,762	28,315	gender skin-tone
PHASE	GCC [43]	crowd-sourcing	region	18,889	35,347	age gender skin-tone ethnicity emotion activity

and mitigation techniques [5, 11, 23, 54] are essential tools for developing fairer models, however, they require demographic attributes, such as gender or skin-tone, to be available. These annotations are currently scarce, and only exist for a few datasets and attributes [60, 61].

In this paper, we contribute to the analysis, evaluation, and mitigation of bias in vision-and-language tasks by annotating six types of demographic[1] and contextual[2] attributes in a large dataset: the Google Conceptual Captions (GCC) [43], which was one of the first automatically crawled datasets with 3.3 million image-caption pairs. We name our annotations PHASE (Perceived Human Annotations for Social Evaluation), and we use them to conduct a comprehensive analysis of the distribution of demographic attributes on the GCC dataset. We complement our findings with experiments on three main vision-and-language tasks: image captioning, text-image embeddings, and text-to-image generation. Overall, we found that a dataset crawled from the internet like GCC presents big unbalances on all the demographic attributes under analysis. Moreover, when compared against the demographic annotations in MSCOCO by Zhao *et al*. [60], GCC has bigger representation gaps in gender and skin-tone. As for the downstream tasks, the three of them show evidence of different performance for different demographic groups.

2. Related work

Bias in vision-and-language Vision-and-language are a set of tasks that deal with data in image and text format. This includes image captioning [50], visual question answering [4], or visual grounding [36]. In terms of societal bias, Burns *et al*. [11] showed that captions in the standard MSCOCO dataset [13] were gender imbalanced, and proposed an equalizer to mitigate the problem. Since then, gender bias has been found not only in image captioning [2, 22, 46, 51] but also in text-to-image search [54], pretrained vision-and-language models [5, 44], multimodal embeddings [40], visual question answering [21], and multimodal datasets [9]. Zhao *et al*. [60] showed that gender is not the only attribute affected by bias; skin-tone also contributes to getting differences in captions. As the problem is far from being solved, tools to study and mitigate models' different demographic representations are essential.

Bias detection datasets Annotations for studying societal bias in vision-and-language tasks are scarce. Without enough data, it is unfeasible to analyze and propose solutions to overcome the problem. Previous work [60, 61] annotated samples from the MSCOCO dataset [30] with the perceived attributes of the people in the images. First, [61] automatically assigned a binary gender class to images by using the gender words from the captions, excluding images whose captions contained multiple genders. Alternatively, [60] annotated gender and skin-tone via crowdsourcing. In this case, the annotations were conducted at the person-level, as opposed to the whole image, allowing images with multiple people to have multiple annotations. To increase diversity in images other than MSCOCO and attributes other than gender and skin-tone, we annotate PHASE from GCC with six attributes, four demographic and two contextual. Image-text datasets with annotations for bias detection are summarised in Table 1.

3. PHASE annotations

Manually annotating demographic attributes from images poses many challenges, as discussed in detail in [3]. The attributes perceived by external observers may not correspond with the real attributes of the annotated person. Moreover, the definition of some attributes, such as the ones related to race or ethnicity, is ambiguous and subjective [20]. Even so, demographic annotations are essential to measuring how imbalanced a dataset or a model's output is. We attempt to mitigate those problems by first, clari-

[1] Demographic attributes: age, gender, skin-tone, ethnicity.
[2] Contextual attributes: emotion, activity.

fying to both annotators and potential users that the annotations do not correspond to real attributes, but perceived ones; and second, mitigating the effects of subjectivity by collecting multiple annotations per sample, using two race-related attributes instead of one, and sharing the attributes of the anonymized annotators for uncovering potential correlations between annotators background and their perception.

3.1. Image source

The GCC dataset [43] contains about 3.3 million samples from the Internet paired with alt-text captions. Originally, images were filtered to remove pornography while captions were post-processed to transform named-entities into hypernyms, *e.g. Harrison Ford* $\rightarrow$ *actor*. Other than that, no filters were applied to remove toxicity or balance the representations. Due to its large size, GCC has been used for pre-training several vision-and-language models, including VilBERT [31], VLBERT [45], Unicoder-VL [28], UNITER [14], OSCAR [29], or ERNIE-VL [59]. This makes it an ideal testbed for studying how the representation of different demographic attributes affects downstream tasks.

3.2. Attributes

Following [60], we annotate people in images via crowd-sourcing. Details on the annotation process are provided in Section 3.3. For each person, we use Amazon Mechanical Turk[3] (AMT) to get four *demographic* and two *contextual* attributes. With up to six attributes per person, our goal is to analyze bias from an intersectional perspective.

Demographic attributes We denote demographic attributes as characteristics of people that are intrinsic to their being and cannot be easily changed. We annotate four attributes with the following categorization:[4]

- **Age** with five classes: *Baby* (0-2 years-old), *Child* (3-14 years-old), *Young adult* (15-29 years-old), *Adult* (30-64 years-old), *Senior* (65 years-old or more).
- **Gender** with two classes: *Man* and *Woman*.
- **Skin-tone** with six types, *Type 1* to *Type 6*, according to the Fitzpatrick scale [18].
- **Ethnicity** with eight classes from the FairFace dataset [26]: *Black*, *East Asian*, *Indian*, *Latino*, *Middle Eastern*, *Southeast Asian*, *White* plus an extra *Other* class.

Contextual attributes We denote contextual attributes as temporary states captured in an image. We annotate two:

- **Emotion** with five classes: *Happiness*, *Sadness*, *Fear*, *Anger*, *Neutral*.
- **Activity** with ten groups adapted from the ActivityNet taxonomy [12] to fit static images: *Helping and Caring*, *Eating*, *Household*, *Dance and Music*, *Personal Care*, *Posing*, *Sports*, *Transportation*, *Work*, and *Other*.

All the attributes also include an *unsure* class.

[3] https://www.mturk.com/

[4] Demographic categorization systems cannot represent all the different identities, and thus, it should only be seen as a rough and non-inclusive approximation to different social groups in order to analyze disparities.

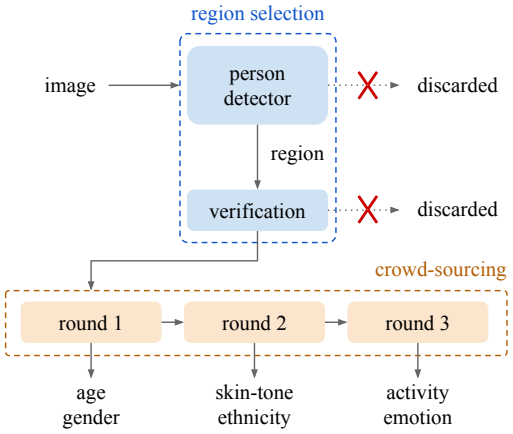

Figure 2. Annotation process. In the region selection part, regions with people are detected and filtered. In the crowd-sourcing part, the selected regions are annotated in three rounds.

3.3. Annotation process

Due to the large size of the GCC dataset, we annotate all the validation set ($4,614$ images with people) and a random subset of the training set ($14,275$ images with people). The annotation process[5] consists of two parts, region selection and crowd-sourcing, summarized in Figure 2.

Region selection Unlike MSCOCO, GCC does not have object region annotations. There are machine-generated labels for a subset of the training images. We do not rely on these labels, as they are not associated with bounding boxes and are not available for the validation set. Instead, we conduct a region selection process semi-automatically in two phases. In the first phase, we run the object detector YOLOv5 [25] to detect regions with objects. For each detected region, YOLOv5 returns an object class, a confidence score, and a bounding box. We keep regions whose object label is *person*, the confidence score is higher than 0.35, and the pixel area is larger than $5,000$. As the detected regions are relatively noisy, in the second phase, we conduct a manual verification. We discard regions without a person, depicting multiple people, or not photographs.

Crowd-sourcing The selected regions are annotated via AMT crowd-sourcing in three rounds. In the first round, we collect the demographic attributes of age and gender. In the second round, the ones related to race, *i.e.* skin-tone and ethnicity. Lastly, we collect the contextual attributes of emotion and activity. Splitting the annotation process into different rounds allows us to have better control of the quality.

[5] Annotation process approved by the Institutional Review Board (IRB).

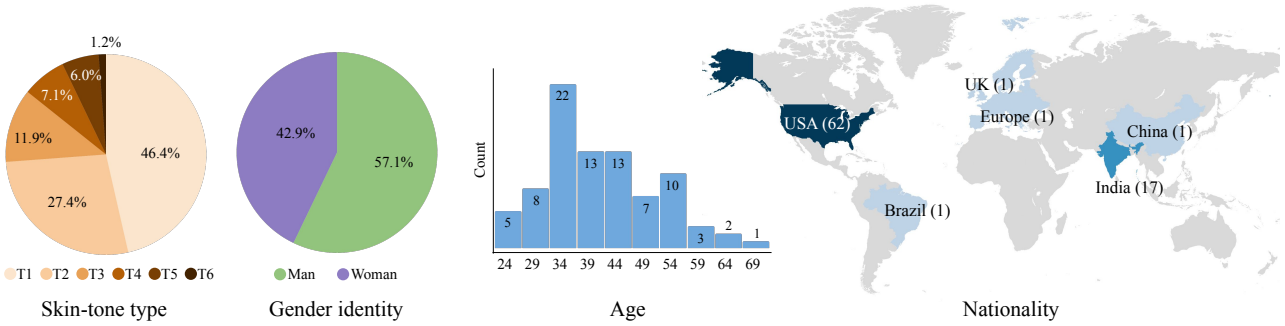

Figure 3. Statistics on workers' self-reported attributes. From left to right: skin-tone type pie chart, gender identity pie chart, age histogram with bin width of 5 years, and nationality map.

For the first two rounds, we expose workers to the cropped regions, to avoid the context being used to predict demographic attributes [6]. For the last round, as the context is essential, we show workers the full image with a bounding box around the person of interest. We ask them to select the perceived attributes from a given list. For all rounds, three different workers annotate each region. To ensure quality, we conduct random verifications of the annotated samples.

Workers must fulfill three conditions: 1) to be *Masters*, which is a qualification granted by the platform to workers with high-quality work, 2) to agree to a consent form as per our Institutional Review Board approval, and 3) to conduct a survey for collecting workers' demographics. As the perception of attributes may be affected by the own workers' attributes [15], we anonymize workers' survey and release it with the rest of the annotations.

4. Annotator statistics

Annotator demographics The workers' survey has the following questions: *age* as an input text box; *gender* as a multiple choice form with options *man*, *woman*, *non-binary*, *others*; *nationality* as an input text box; and *skin-tone type* as a multiple choice form with six options according to the Fitzpatrick scale. Statistics are shown in Figure 3. In total, 84 workers answered the survey. From these, 9.5% reported to be less then 30 year-old, 88.1% between 30 and 64 year-old, and 2.4% more than 65 year-old. With respect to gender, 42.9% workers reported to be women, and 57.1% men, with no other genders reported. Skin-tone is predominantly light (Types 1, 2, and 3), with only 14.3% skin-tone Types 4, 5, or 6. Most of the workers are from the United States of America (73.8%) followed by India (20.2%). Other nationalities include Brazil, the United Kingdom, Europe, and China, with a worker each.

Inter-annotator agreement We measure to what extent attributes are equally perceived by different people by computing the inter-annotator agreement. We use two metrics: a consensus ratio and the Fleiss' kappa [19]. Consensus ratio, $n+$, indicates the percentage of regions in which n or more workers give the same class, and Fleiss' kappa, κ, measures whether the annotations agree with a probability above chance. Results are in Table 2. According to κ, gender has an almost perfect agreement, with 92.5% of the regions with a consensus of three workers. Activity shows a substantial agreement, whereas age and ethnicity have a moderate agreement. Considering that ethnicity is known to be a subjective attribute [20], the relatively high agreement may be due to workers having a similar background (*e.g.* most workers are from the United States of America). In contrast, skin-tone and emotion have the lowest κ (0.24 and 0.37, respectively) but are still well above chance ($\kappa \leq 0$). Note that skin-tone can be affected by image illumination and the annotator's own perception of color, making consecutive skin-tone types (*e.g. Type 2* and *Type 3*) difficult to distinguish. Thus, we additionally check agreement in binary skin-tone classification, *i.e.* lighter skin-tone (*Types 1, 2, 3*) and darker skin-tone (*Types 4, 5, 6*). κ increases from 0.24 (fair) to 0.59 (moderate), so from now on we preferably use binary skin-tone unless otherwise stated.

Table 2. Inter-annotator agreement in the training set. 3+ and 2+ indicate the ratio (%) of regions with a consensus of three or more, or two or more workers, respectively. κ indicates Fleiss' kappa and Agreement is set according to κ as in [49]. Labels indicates the number of classes plus the *unsure* class (+1).

Attribute	Labels	3+	2+	κ	Agreement
age	$5+1$	48.8	97.2	0.44	moderate
gender	$2+1$	92.5	99.6	0.89	almost perfect
skin-tone	$6+1$	27.1	86.3	0.24	fair
skin-tone (binary)	$2+1$	80.0	99.4	0.59	moderate
ethnicity	$8+1$	50.9	90.1	0.41	moderate
emotion	$5+1$	46.5	94.2	0.37	fair
activity	$11+1$	61.8	95.3	0.65	substantial

5. PHASE analysis

We annotated the whole validation set and a random portion of the training set. We downloaded 13,501 and 2,099,769 validation and training images, respectively; from those, 5,668 and 498,006 validation and training images had detected human regions by YOLOv5; from those, all the validation (5,668) and a random subset of the training images (17,147), were annotated; only 4,614 and 14,275 validation and training images passed the manual verification and their human-detected regions were annotated with the six demographic and contextual attributes. Overall, 35,347 regions were annotated: 8,833 in the validation set and 26,514 in the training set.

We publicly share the data in two formats: *raw*, in which each region has up to three annotations, and *region-level*, in which each region is assigned to a class by majority vote. If there is no consensus, the region is labeled as *disagreement*.

Attribute analysis Full statistics per attribute and class are reported in the supplementary material. Overall, all the attributes are imbalanced, with one or two predominant classes per attribute. For age, the predominant class is *adult*, appearing in 45.6% of the region-level annotations, while for gender, the gap between *man* (63.6%) and *woman* (35.1%) is 28.5 points. The gap is bigger in skin-tone and ethnicity: skin-tone *Type 2* is annotated in 47.1% of the regions, while *Types 4*, *5*, and *6* together only in 17.5%. Similarly, in ethnicity, *White* class is over-represented with 62.5% regions, while the rest of the classes appear from 0.6% to 10% regions each. When conducting an intersectional analysis, big differences arise. For example, the classes *man* and *White* appear together in 13,651 regions, whereas *woman* and *Black* in only 823.

As per the contextual attributes, the most represented emotion is *neutral* (47.1%) followed by *happy* (35.7%), whereas negative-associated emotions (*sad*, *fear*, *anger*) are just 2.33% in total. For activity, the most common classes are *posing* (28.6%) and *other* (20.5%).

Gender and context We cross-check gender with contextual attributes. Region-level emotion statistics per gender are shown in Figure 4, where it can be seen that women tend to appear *Happy* whereas men tend to appear *Neutral*. This aligns with the gender stereotyping of emotions, which is a well-documented phenomenon in psychology [35]. We also detect disparities in activities per gender, especially in the classes *posing* and *sports*: in *woman* regions, there are 42.8% *posing* and 5.1% *sports* annotations, while in *man* regions, 21% are *posing* and 26.9% are *sports*.

GCC vs MSCOCO We compare gender and skin-tone annotations in GCC against the annotations in MSCOCO by Zhao *et al*. [60]. In MSCOCO, gender was reported as 47.4% *man* regions and 23.7% *woman* regions, which is a gap of 23.7 points, smaller than the 28.5 points gap in GCC. With respect to skin-tone, the gap in MSCOCO between lighter skin-tone (*Types 1*, *2*, and *3*) and darker skin-tone (*Types 4*, *5*, and *6*) was reported to be 52.9 (from 61% to 8.1%), whereas the gap in GCC is 64 (from 79.7% to 15.7%). This shows that GCC, an automatically crawled dataset, is more unbalanced than MSCOCO, a manually annotated dataset, both in gender and skin-tone attributes.

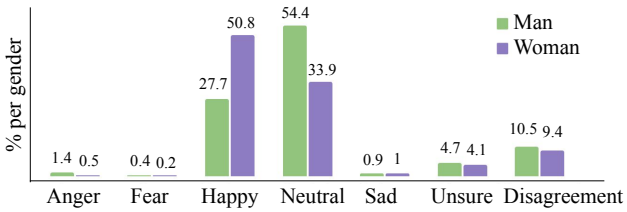

Figure 4. Region-level emotion statistics per gender (percentage).

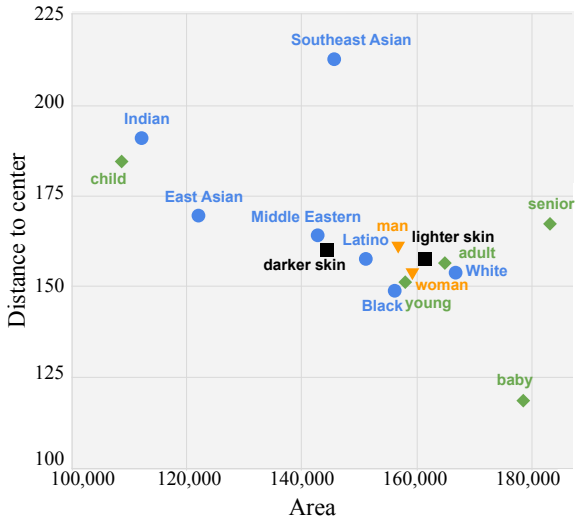

Figure 5. Average area and distance to the image center (in pixels) for each demographic class in the annotated GCC dataset. Colors and shapes indicate classes in the same attribute.

Region analysis We analyze how the regions of the different demographic classes are presented in the GCC dataset. In particular, in Figure 5, we show an average of the area and the distance to the image center per class. The most noticeable disparities are in the attribute age, with *senior* and *baby* being the larger regions from all the classes and *child* the smaller. The class *baby* also tends to be more centered than people in other ages. In contrast, gender regions do not present big differences in size and position. With respect to skin-tone and ethnicity, the area for *lighter* class is bigger than for *darker* skin-tone, whereas ethnicity presents a large variance, with the *Southeast Asian* class being the most faraway from the center and the *White* class the largest one.

Table 3. Image captioning models in terms of bias metrics on PHASE .

	Age		Gender			Skin-tone (binary)		Ethnicity	
	LIC_M	LIC	LIC_M	LIC	Error	LIC_M	LIC	LIC_M	LIC
Unbiased	*25.0*	*0.0*	*50.0*	*0.0*	*0.0*	*50.0*	*0.0*	*14.3*	*0.0*
OFA [55]	61.1 ± 3.7	7.9	74.7 ± 3.7	2.0	4.3	60.6 ± 4.9	-2.4	22.6 ± 5.4	0.9
ClipCap [33]	57.4 ± 3.7	6.2	76.7 ± 1.8	2.7	7.9	68.0 ± 6.5	6.2	24.8 ± 7.5	2.9

6. Downstream bias evaluation

With PHASE annotations we can evaluate societal bias in vision-and-language tasks. We explore image captioning, text-image embeddings, and text-to-image generation.

Image annotations The annotations in PHASE correspond to regions, but the three tasks analyzed in this section use the full image as a whole. To deal with this discrepancy, we transform the region-level annotations into image-level annotations. To that end, we only use images in which all the region-level annotations belong to the same class. For example, an image is labeled as *woman* if all the region-level annotations of gender are *woman*. Images containing different classes for a specific attribute are not used. Note that this approach is not unique and other methods can be used according to the type of evaluation to be conducted. In addition, to deal with the low number of annotations, we merge *baby* and *child* into a single class *baby & child*.

6.1. Image captioning

Image captioning is one of the reference tasks in vision-and-language research. Given an image, a captioning model generates a sentence describing its contents. Multiple image captioning models trained on the MSCOCO dataset have been shown to be biased with respect to gender [11] and skin-tone [60]. In this section, we evaluate two of the latest models trained on GCC, OFA [55] and ClipCap [33], in the four demographic attributes in our annotations. While OFA is a state-of-the-art model based on vision-and-language Transformers [48], ClipCap leverages CLIP embeddings [37], which are also analyzed in terms of bias in the next section.

Metrics We evaluate image captioning models on societal bias and bias amplification with LIC_M and LIC metrics [22], respectively. LIC_M corresponds to the accuracy of a caption classifier that predicts the class of a demographic attribute after masking class-revealing words.[6][7] If the caption classifier is more accurate than random chance, it means that captions of people from different classes are semantically different. The classifier is trained 10 times with different random seeds, and results are reported as average and standard deviation. On the other hand, $LIC = LIC_M - LIC_D$ measures bias amplification by comparing the accuracy of the caption classifier trained on the human captions, LIC_D, against the model-generated captions, LIC_M. If $LIC > 0$, the generated captions are more biased than the original ones and the model amplifies the bias.

Additionally, for the gender attribute, we compute Error [11], which measures the percentage of captions in which gender is misclassified. Error can only be computed if attributes are explicitly mentioned in the generated captions, which is the case for gender, but not for *e.g.* skin-tone.

Results Results are shown in Table 3. Both OFA and ClipCap have a LIC_M score well above the unbiased case for the four attributes. When trained on GCC, except OFA in skin-tone, both models amplify bias with respect to the original dataset. This highlights the urgency of incorporating bias mitigation techniques, such as the model-agnostic method in [23]. Although age is an attribute that is not often analyzed, results show that it is the one with the highest bias, both on LIC_M and LIC metrics. This highlights the urgency to consider age in representation fairness. Results on gender and skin-tone also reveal important biases in the models' outputs, while the large standard deviation in ethnicity, which may be due to the higher number of classes (7) and the smaller number of samples per class (64), makes it difficult to extract reliable conclusions.

6.2. Text-image CLIP embeddings

Next, we evaluate the performance of pre-trained text-image CLIP embeddings [37] for the different demographic attributes. CLIP is a dual architecture with a text encoder and an image encoder that learns image and text embeddings by predicting matching pairs. Due to the large amount of data used for training (400 million image-text pairs), the two encoders learn the correspondences of high-level semantics in the language and the visual modalities. The goal of our evaluation is to check whether the demographic attributes of people in the images have an impact on the accuracy of the embeddings. To conduct this evaluation, we ex-

[6] Gender-revealing words: actor, actress, aunt, boy, boyfriend, brother, chairman, chairwoman, cowboy, daughter, dude, emperor, father, female, gentleman, girl, girlfriend, guy, he, her, hers, herself, him, himself, his, husband, lady, male, man, mother, policeman, policewoman, pregnant, prince, princess, queen, king, she, sister, son, uncle, waiter, waitress, wife, woman (and their plurals).

[7] Age-revealing words: adult, aged, baby, child, elderly, infant, kid, teenager, toddler, young (and their plurals).

Table 4. CLIP embeddings evaluation on PHASE validation set.

Attribute		Samples	R@1	R@5	R@10
age	baby & child	350	44.0	65.4	74.0
	young	1,349	30.4	51.3	60.9
	adult	1,509	27.3	46.7	55.9
	senior	128	44.5	64.1	71.1
gender	man	1,950	32.0	53.2	63.1
	woman	1,617	30.6	49.8	59.1
skin-tone	lighter	3,166	30.2	50.6	59.9
	darker	318	31.1	54.1	62.3
ethnicity	Black	194	29.4	51.5	58.8
	East Asian	58	34.8	56.9	63.8
	Indian	90	34.4	61.1	68.9
	Latino	28	21.4	39.3	50.0
	Middle Eastern	16	31.3	62.5	75.0
	Southeast Asian	16	31.3	37.5	56.3
	White	2,231	30.6	50.6	59.5

tract image and text embeddings with the pre-trained CLIP encoders. For each of the $4,614$ images in the validation set, we rank the validation captions according to the cosine similarity between their embeddings and analyze the accuracy of the ranked list.

Metrics We evaluate accuracy of image-text embeddings as recall at k (R@k) with $k = \{1, 5, 10\}$. R@k indicates the percentage of images whose matching caption is ranked within the top-k positions. We compare the difference in R@k for classes in the same attribute to check whether CLIP embeddings perform differently. In the ideal scenario of unbiased representations, R@k performance in different class attributes should be alike.

Results Results are shown in Table 4. Within each of the four demographic attributes, noticeable differences in performance can be observed. Note that the number of samples in each class is different, which could influence the results. In the supplementary material, we report an evaluation when using the same number of samples per class, in which we verify that the conclusions are not affected. We summarize the main findings as follows:

- For the age attribute, *baby & child* and *senior* have the best R@k, while *young* and *adult* fall well behind with a difference of up to 18.1 in R@10. The big differences in the age classes are consistent with the results of the region analysis (Section 5) and the image captioning (Section 6.1), being the attribute with the highest class variance.
- In gender, *man* samples perform consistently better than *woman* samples.
- For skin-tone, *darker* has higher R@k than *lighter*. This may be explained by the language bias, a documented phenomenon in which skin-tone descriptions are usually omitted for *lighter* but not for *darker* skin-tones [34,47,60].
- Ethnicity is not consistent. For R@1, the highest classes are *Eastern Asian* and *Indian*, but for R@5 and R@10 is *Middle Eastern*. The lowest classes are *Latino*, *Southeast Asian*, and *Black*. The lack of consistency, together with the low inter-annotator agreement (Section 4), shows that ethnicity annotations are highly subjective.
- More samples do not ensure a better recall, which means that the number of samples is not the (only) source for difference in performance. For example, in age, although *adult* and *young* classes are the most common, their performance is the worst. The same happens in skin-tone; despite the predominance of *lighter*, R@k is higher for *darker* samples. In contrast, in gender, *man* outperforms *woman* both in the number of samples and in R@k.

6.3. Text-to-image generation

Lastly, we analyze the demographic representation on Stable Diffusion [39], one of the latest text-to-image generation models. Text-to-image generation, which can be seen as the reverse operation of image captioning, consists on creating an image from a text sentence, also known as *prompt*. In particular, Stable Diffusion relies on pre-trained CLIP embeddings and Diffusion Models [24] to generate an image in the latent space whose embedding is close to the input prompt embedding. In our evaluation, we use the $4,614$ captions in the validation set as prompts to generate an image per caption. We use the demographic annotations of the original images associated with the captions to study Stable Diffusion representations.

Metrics The official code for Stable Diffusion v1.4[8] includes a Safety Checker module that raises a flag when the generated images are considered to be NSFW.[9] The module is pre-trained and used off-the-shelf by the community. We check whether there are patterns in the output of the Safety Checker according to the demographic attributes of the input caption. Additionally, we compare the demographics of the generated images against the demographics of the original images associated with the captions.

Results Out of $4,614$ generated images, 36 are flagged as unsafe by the Safety Checker module. From these, we do not find prominent differences between the distribution of classes in the original images and the unsafe images for age, skin-tone, and ethnicity attributes. We do find, however, the distribution of gender unusual: despite *woman* only being 35.04% validation images, it raises 51.61% unsafe images.

[8] https://github.com/CompVis/stable-diffusion

[9] Not safe for work, a tag commonly used for pornographic, violent, or otherwise inappropriate content.

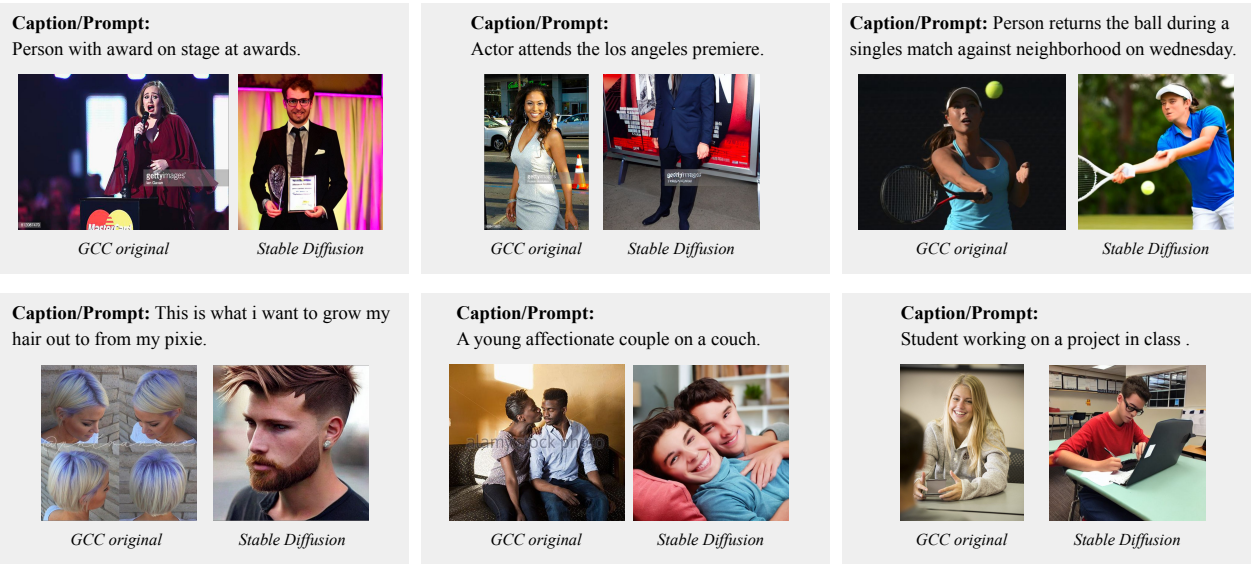

Figure 6. Examples of Stable Diffusion [39] generated images with the caption used as a prompt, together with the original GCC image. Stable Diffusion tends to generate images of White men for captions with neutral language (*e.g.* person, student, couple, *etc.*).

This indicates that gender, especially *woman*, has an important contribution to the Safety Checker. With the current experiments, we cannot clarify whether this is due to Stable Diffusion being more prone to generate NSFW images from prompts from women images, or due to the Safety Checker being more prone to detect NSFW in woman-generated images. A table with the results can be found in the supplementary material. Additionally, in Figure 6, we show some examples of Stable Diffusion generated images and compare them against the original images in GCC. We observe that when the prompt refers to people in a neutral language (*e.g.* person), the generated images tend to represent *White men*. To verify this observation, we manually annotate 100 generated images. For neutral language prompts, we observe 47.4% men *vs.* 35.5% women; and 54.0% lighter *vs.* 26.6% darker skin-tone. These results are consistent with concurrent work on text-to-image generation bias [7].

7. Limitations

Although we argued that demographic annotations are necessary to address societal bias in vision-and-language models, we acknowledge that they pose some risks.

Perceived attributes The annotations are conducted by external observers, meaning they reflect perceived attributes. Perceived attributes may not correspond to the real person's attributes. As per dataset construction, GCC in this case but computer vision datasets in general, it is not possible to ask people depicted in the images about their self-perceived attributes. Annotations should not be considered real, objective, or trustworthy labels, but an approximation of how observers classify people in images.

Subjectivity The annotations are subjective and not universal. Many demographic attributes, especially the ones related to race, ethnicity, or emotion, have different classification systems according to different contexts and cultures.

Malicious uses The intended use of the annotations is for research on societal bias and fairness. Although we cannot control who, when, and how will use the annotations once they are publicly available, the use for malicious applications is strictly prohibited.

8. Conclusion

We studied social bias in large and uncurated vision-and-language datasets. Specifically, we annotated part of the GCC dataset with four demographic and two contextual attributes. With annotations on age, gender, skin-tone, ethnicity, emotion, and activity, we conducted a comprehensive analysis of the representation diversity of the dataset. We found all six attributes to be highly unbalanced. When compared against manually annotated datasets such as MSCOCO, GCC presented bigger gaps in gender and skin-tone, with an overrepresentation of *man* and *lighter* skin-tones. Additionally, we evaluated three downstream tasks: image captioning, image-text CLIP embeddings, and text-to-image generation. In all the tasks, we found differences in performance for images in different demographic classes, highlighting the need for resources and solutions.[10]

[10]This work is partly supported by JST CREST Grant No. JPMJCR20D3, JST FOREST Grant No. JPMJFR216O, JSPS KAKENHI No. JP22K12091, and Grant-in-Aid for Scientific Research (A).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1

[2] Jack J Amend, Albatool Wazzan, and Richard Souvenir. Evaluating gender-neutral training data for automated image captioning. In *International Conference on Big Data*, 2021. 2

[3] Jerone TA Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, Shruti Nagpal, and Alice Xiang. Ethical considerations for collecting human-centric image datasets. *arXiv preprint arXiv:2302.03629*, 2023. 2

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 2

[5] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *AACL-IJCNLP*, 2022. 2

[6] Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*, 2019. 4

[7] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*, 2022. 8

[8] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *WACV*, 2021. 1

[9] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 1, 2

[10] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, 2018. 1

[11] Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 2, 6

[12] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 3

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 3

[15] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Trans. ACL*, 2022. 4

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[17] Karan Desai, Gaurav Kaul, Zubin Trivadi Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks Track*, 2021. 1

[18] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology*, 124(6):869–871, 1988. 3

[19] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378 – 382, 1971. 4

[20] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *FAccT*, 2020. 2, 4

[21] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Gender and racial bias in visual question answering datasets. In *FAccT*, 2022. 1, 2

[22] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *CVPR*, 2022. 1, 2, 6

[23] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Model-agnostic gender debiased image captioning. In *CVPR*, 2023. 2, 6

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 7

[25] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Kalen Michael, Jiacong Fang, imyhxy, Lorna, Colin Wong, (Zeng Yifu), Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Max Strobel, Mrinal Jain, Lorenzo Mammana, and xylieong. ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations, Aug. 2022. 3

[26] Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, 2021. 3

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Trans. IJCV*, 123(1):32–73, 2017. 1

[28] Gen Li, N. Duan, Yuejian Fang, Daxin Jiang, and M. Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 3

[29] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 3

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 2

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019. 3

[32] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. *arXiv preprint arXiv:2206.09191*, 2022. 1

[33] Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 6

[34] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. How do we talk about other people? Group (un) fairness in natural language image descriptions. In *AAAI HCOMP*, 2019. 7

[35] E Ashby Plant, Janet Shibley Hyde, Dacher Keltner, and Patricia G Devine. The gender stereotyping of emotions. *Psychology of women quarterly*, 24(1):81–92, 2000. 5

[36] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 6

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 7, 8

[40] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *NAACL*, 2021. 1, 2

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1

[42] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop Datacentric AI*, 2021. 1

[43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1, 2, 3

[44] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Workshop on Gender Bias in Natural Language Processing*, 2022. 2

[45] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. 3

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

[47] Emiel Van Miltenburg. Stereotyping and bias in the Flickr30k dataset. In *Workshop on Multimodal Corpora*, 2016. 7

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 6

[49] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363, 2005. 4

[50] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2

[51] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring representational harms in image captioning. In *FAccT*, 2022. 2

[52] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. In *ECCV*, 2020. 1

[53] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *ICML*, 2021. 1

[54] Jialu Wang, Yang Liu, and Xin Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *EMNLP*, 2021. 2

[55] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 6

[56] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 2019. 1

[57] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *FAccT*, 2022. 1

[58] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *FAccT*, 2020. 1

[59] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *AAAI*, 2021. 3

[60] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, 2021. 1, 2, 3, 5, 6, 7

[61] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017. 1, 2