Uncurated Image-Text Datasets: Shedding Light on Demographic Bias Supplementary Material

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

This document includes additional information to support the claims and experimental results in the main paper. The document is divided into the following sections:

- Section A: PHASE statistics.
- Section B: YOLOv5 bias evaluaiton.
- Section C: CLIP evaluation results.
- Section D: Stable Diffusion results.

A. PHASE statistics

Statistics about the number of annotations per attribute and class are reported in Table 3. For each class, we show the *number of annotations* as the raw annotations for the given class, the *number of regions* as the region-level annotations reached after annotator majority voting, and the *number of images* as the images with at least one region-level annotation with the class of interest.

B. YOLOv5 bias evaluation

Following findings about bias in pre-trained object detection models [1], we evaluate YOLOv5 in terms of skin-tone bias to check whether the use of this model can be a contributing factor to the representation discrepancies in PHASE annotations. We detect people in MSCOCO and compare accuracy per skin-tone using [2] annotations. Results are reported as follows: *darker* skin-tone recall is 0.49, and *lighter* skin-tone recall is 0.55. This shows that there is, indeed, a difference in performance according to skin-tone. However, we believe that it is not as big as to justify the representation gap that was found in the GCC dataset and the conclusions of our analysis still stand.

C. CLIP evaluation results

We report the results of the CLIP evaluation when balancing the number of samples per class and attribute. Table 4 compares CLIP performance in R@k for $k = 1, 5, 10$ when using all the samples in the validation set (Unbalanced), and when using the same number of samples per class and attribute (Balanced). For the balanced results, we use the number of samples in the smallest class in each attribute. Results are reported as mean and standard deviation over 100 runs with different random samples per class.

Table 1. CLIP embeddings evaluation on PHASE validation set when detected person bounding boxes are occluded (black).

Attribute		Samples	R@1	R@5	R@10
age	baby & child	350	12.9	23.4	29.1
	young	1,349	8.6	16.7	21.9
	adult	1,509	9.4	19.9	25.6
	senior	128	11.7	28.9	34.4
gender	man	1,950	12.0	24.3	30.6
	woman	1,617	8.1	15.3	20.0
skin-tone	lighter	3,166	8.5	17.5	22.9
	darker	318	11.6	25.2	29.6
ethnicity	Black	194	9.3	20.1	23.2
	East Asian	58	8.6	31.0	36.2
	Indian	90	11.1	28.9	35.6
	Latino	28	7.1	14.3	17.9
	Middle Eastern	16	12.5	12.5	25.0
	Southeast Asian	16	12.5	12.5	12.5
	White	2,231	8.7	17.4	23.0

CLIP with occlusions To better understand what about the image leads to differing performances, we occlude the detected bounding boxes and repeat the evaluation process. We find that masking people bounding boxes makes CLIP's R@1 drop from about 30% to 8% for all the attributes, which means that relevant information is contained in person regions. Moreover, as shown in Table 1, the conclusions are maintained, *e.g.* recall for *man* is higher than for *woman*, which suggests that part of the bias is from the language.

D. Stable Diffusion results

Table 2 reports the statistics of Stable Diffusion's Safety Checker per attribute and class. We compare the percentage

Table 2. Classes in the validation set and classes labeled as unsafe by Stable Diffusion's Safety Checker.

Attribute	Validation set (%)	Unsafe label (%)
age		
baby	0.89	3.23
child	6.70	12.90
young	29.24	32.26
adult	32.70	32.26
senior	2.77	3.23
unsure	*1.11*	*0.00*
multiple	*26.59*	*16.13*
gender		
man	42.26	35.48
woman	35.04	51.61
unsure	*0.98*	0.00
multiple	*21.72*	12.90
skin-tone (binary)		
lighter	68.62	80.65
darker	6.89	3.23
unsure	*5.66*	*3.23*
multiple	*18.83*	*12.90*
ethnicity		
Black	4.20	3.23
East Asian	1.26	0.00
Indian	1.95	3.23
Latino	0.61	0.00
Middle Eastern	0.35	0.00
Southeast Asian	0.35	0.00
White	48.35	64.52
unsure	*5.52*	*3.23*
multiple	*37.41*	*25.81*

of samples per class in the image annotations, with the percentage of samples per class labeled as unsafe by the Safety Checker. It stands out that the class *woman* raises 51.61% of the unsafe labels whereas only accounts for 35.04% of the original images. *Lighter* skin-tone and *White* ethnicity also show big increases in the percentage of samples raised as unsafe, but differently from the *woman* class, both of them are the predominant class in their respective attributes.

References

[1] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019. 1

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

Table 3. Statistics of annotations in PHASE per attribute and class. *Annotations* reports the raw number of annotations per class. *Regions* is the total number of region-level annotations after majority voting. *Images* accounts for the number of images with at least one region-level annotation with the class. Due to the inter-annotator agreement results, skin-tone region-level annotations are conducted for binary skin-tone only. For each attribute, the most common class is highlighted in **bold** and the unsure class in *italics*.

Attribute	Annotations	Regions	Images	Attribute	Annotations	Regions	Images
age	106,041	35,347	18,889	skin-tone type	105,801	-	-
baby	955	306	259	type 1	15,388	-	-
child	7,829	2,578	1,569	**type 2**	**49,821**	-	-
young adult	40,398	13,313	8,841	type 3	18,083	-	-
adult	**48,604**	**16,117**	**10,631**	type 4	8,219	-	-
senior	4,632	1,375	1,152	type 5	5,771	-	-
unsure	*3,623*	*653*	*525*	type 6	4,570	-	-
gender	106,041	35,347	18,889	*unsure*	*3,949*	-	-
man	**67,122**	**22,491**	**13,511**	skin-tone (binary)	-	35,347	18,889
woman	36,936	12,406	8,329	**lighter**	-	**28,187**	**16,245**
unsure	*1,983*	*285*	*241*	darker	-	5,572	3,838
ethnicity	105,801	35,347	18,889	*unsure*	-	*922*	*730*
Black	11,314	3,664	2,657	activity	97,021	35,347	18,889
East Asian	3,957	953	707	caring	786	127	119
Indian	4,434	980	616	music	14,706	5,012	3,218
Latino	8,826	1,309	1,168	eating	1,012	278	206
Middle Eastern	5,513	373	349	household	180	19	18
Southeast Asian	2,706	211	174	personal	291	39	33
White	**63,253**	**22,098**	**13,698**	**posing**	**30,409**	**10,121**	**6,619**
unsure	*5,578*	*1,289*	*1,021*	sports	19,933	6,725	3,807
emotion	100,248	35,347	18,889	transportation	811	224	181
happy	41,059	12,603	8,215	work	5,043	1,433	891
sad	2,221	331	308	sports	19,933	6,725	3,807
fear	1,205	117	114	other	22,770	7,249	4,247
anger	2,391	377	346	*unsure*	*1,080*	*164*	*149*
neutral	**47,367**	**16,646**	**10,473**				
unsure	*6,005*	*1,663*	*1,224*				

Table 4. CLIP evaluation on PHASE validation set. Results are reported as R@k for $k = 1, 5, 10$. *Unbalanced* denotes when all the samples are used, resulting in highly unbalanced classes. *Balanced* denotes when using the same number of samples per class and attribute.

		Unbalanced				Balanced			
Attribute	Class	Size	R@1	R@5	R@10	Size	R@1	R@5	R@10
age	baby & child	350	44.0	65.4	74.0	128	44.0 ± 3.5	65.3 ± 3.4	73.9 ± 3.1
	young	1,349	30.4	51.3	60.9	128	29.8 ± 3.9	51.0 ± 4.5	60.8 ± 4.4
	adult	1,509	27.3	46.7	55.9	128	27.5 ± 3.8	46.5 ± 4.2	55.4 ± 4.1
	senior	128	44.5	64.1	71.1	128	44.5 ± 0.0	64.1 ± 0.0	71.1 ± 0.0
gender	man	1,950	32.0	53.2	63.1	1,617	32.1 ± 0.4	53.2 ± 0.5	63.1 ± 0.4
	woman	1,617	30.6	49.8	59.1	1,617	30.6 ± 0.0	49.8 ± 0.0	59.1 ± 0.0
skin-tone	lighter	3,166	30.2	50.6	59.9	318	30.1 ± 2.6	50.4 ± 2.8	60.1 ± 2.8
	darker	318	31.1	54.1	62.3	318	31.1 ± 0.0	54.1 ± 0.0	62.3 ± 0.0
ethnicity	Black	194	29.4	51.5	58.8	16	29.1 ± 11.5	49.9 ± 11.2	57.7 ± 11.4
	East Asian	58	34.8	56.9	63.8	16	33.5 ± 10.5	56.4 ± 10.5	64.2 ± 10.7
	Indian	90	34.4	61.1	68.9	16	34.4 ± 12.3	61.9 ± 11.8	69.5 ± 11.4
	Latino	28	21.4	39.3	50.0	16	21.4 ± 6.2	37.9 ± 8.2	48.6 ± 8.6
	Middle Eastern	16	31.3	62.5	75.0	16	31.3 ± 0.0	62.5 ± 0.0	75.0 ± 0.0
	Southeast Asian	16	31.3	37.5	56.3	16	31.3 ± 0.0	37.5 ± 0.0	56.3 ± 0.0
	White	2,231	30.6	50.6	59.5	16	31.2 ± 12.6	50.8 ± 12.3	59.3 ± 11.7