A. Appendix

A.1. Category Selection

Imageability and Salience of the Categories When choosing the categories for FACET , we considered the "imageability" of our concepts, from [96]. However, we found that this did not transfer well to our our use case. First, we found that many of the 'highly imageable' concepts include classes directly related to a demographic attribute. *E.g. black woman (n09637339) has an imageability score of 5 out of 5.* Additionally, many highly imageable concepts are abstract, meaning they are easy to imagine but hard classify. E.g. It is easy to imagine what the concept `mother` may look like, but it is hard to determine if someone is a "mother" from a photo. *Is any person perceived to be feminine presenting with a child in a photo presumed to be a mother?*)

Class Hierarchy and Representation We show the full connection of our chosen concepts in WordNet in their relation to the `Person` synset. Figure 8 shows the full connection of our chosen concepts in WordNet in their relation to the `Person` synset. All relevant sub-trees and intermediate synsets are shown. We can see that many of the classes in FACET share the same parent node. We also note that no class in FACET is a direct descendant of another class. This demonstrates that there is no overlap between classes. Table 17 shows the representation of each class in the evaluation set.

A.2. Annotation Pipeline Design

We descibe in more detail the annotation pipeline we use for FACET.

A.2.1 Annotation Pipeline Design

Preprocessing Figure 6 shows the pre-processing steps of the captions to create the candidate set of images to annotate. We 'score' each caption for each category based on the overlap of relevant words for the category and caption. We sample captions with the highest 'score' per category. We choose the candidate images for FACET from a set of roughly 6 million images.

We select a starting set of images for annotation such that we expect the portion of images that pass stage 1 to be roughly class balanced. To approximate the probability that images with overlap per category are true positives, we sample 50 images per category and annotate the true positives. We use this frequency to determine how much to over sample a specific category. As we continue the annotation process, for additional rounds, we sample images with overlap based on the categories that are under-represented in the dataset thus far. We note that many categories did not have enough images with matching relevant words and as such we did not achieve equal representation of all categories.

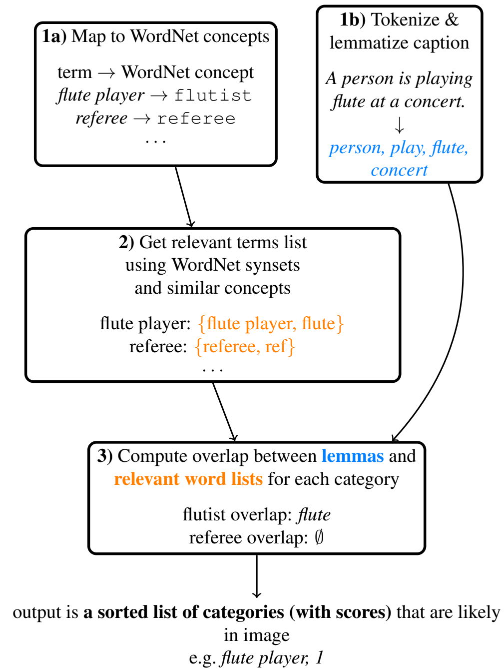


Figure 6: **Label annotation pipeline:** The preprocessing steps before beginning the annotation pipeline. **In 1a)** we map all of the person-related classes to concepts in WordNet. We denote WordNet concepts in a different font. (See Section 3 for a full description on WordNet concepts and synsets). **In 1b)** we tokenize and lemmatize the captions to produce a list of **lemmas**. **In 2)** for each of the 52 categories, we generate a **relevant words list** using WordNet synsets. Lastly **in 3)** we compute overlap between the **lemmas** and **relevant words list** and select images to annotate which have high overlap.

Annotation Stages Figure 7 shows the four separate annotation tasks of the main annotation pipeline. Breaking the annotation process into multiple sub-tasks allows for more fine-grained control. For stage 1, we focus on speed, and ask annotators to spend little time per task. To increase speed, we group multiple images with the same target categories into the a single task with a default value of *0 people match the categories*, and ask the annotator to label each image. We separate stages 3 and 4, so that we can gather multiple annotations for apparent skin tone only. We separate these stages from stage 2 to simplify the task for annotators, such that they only need consider the perceived demographic attributes for one person at a time. Additionally, this allows the annotators in later stages to QA the annotations from earlier stages, as described in Section A.3.2.

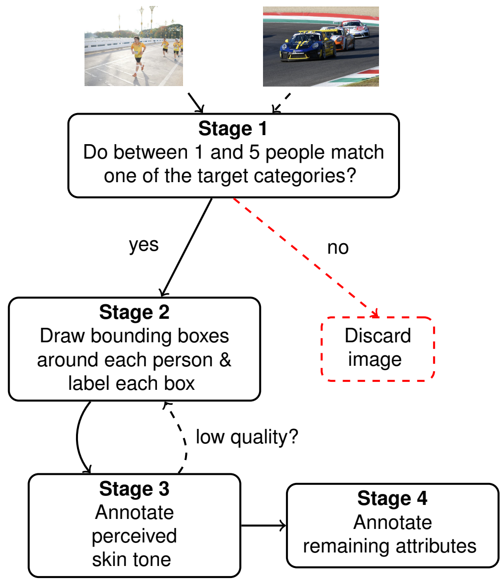


Figure 7: **Image annotation pipeline:** The four stages of the main annotation pipeline. The image on the left can be fully annotated; the image on the right does not contain the target categories and gets excluded after Stage 1. Dashed lines show paths that do not advance to the next stage.

Mask Annotations We collect labels for SA-1Bmask annotations separately after completing the annotation pipeline. First, we select candidate people from FACET with attempts made to balance the number of people per demographic group. Next, we select a candidate set of masks to annotate by collecting the set of masks inside the bounding boxes for these people. For each mask and FACET bounding box in which it resides, we asked annotators if the mask corresponds to *the person's body, the person's hair, an item of clothing on the person* for a given person, as denoted by a bounding box. Annotators did not make any modifications to the masks, e.g. change the shape. Annotators were told to only select a class if the mask covers the entire item; masks for a portion of the person, or part of item of clothing were not labeled. Additionally, annotators were only told to select a class if it met the label for the person described by the bounding box; masks for people, hair, clothing inside of the bounding box but belonging to a different person were not labeled. Thus, each mask is attached to a specific person in FACET. The breakdown of the masks per image is is given in A.4.2. The breakdown of masks per given demographic or additional attribute is given in Table 18.

A.3. Annotation Quality Assurance

A.3.1 Annotator Quality Assurance using Training

Before completing any annotations used in FACET, annotators were trained for each stage separately. We trained annotators by giving them a sample set of tasks and comparing their annotations to a known golden set. For Stages 1, 3 and 4 (image filtering, perceived skin tone annotation, other perceived attributes annotation), annotators passed the training step if the recall of their annotations compared to a fixed golden set was above a quality threshold. This threshold was set for each stage depending on the difficulty of the task. For Stage 2 (drawing bounding boxes) QA was done per annotator to assess the quality of boxes. We provided feedback to annotators individually and only graduated the annotators once they addressed the feedback. A manual IoU threshold of 0.85 between an annotator and the golden set was used. Annotators under that threshold were not manually reviewed, as we found that this correlated with extremely poor box quality, and these annotators did not graduate training. Before feedback, we noticed that many annotators were drawing bounding boxes that included objects the person was holding (*e.g. guitar*) as opposed to tightly around the person. After manual review and feedback, the quality of the annotations was much higher and consistent.

A.3.2 Annotation Quality Assurance using Multi-Review and Quality Checks

In addition to implementing a multi-review process for the perceived skin tone annotations of each target person as discussed in Section 5.2, we used Stage 3 to QA the bounding boxes drawn by the annotators. The annotators in Stage 3 were asked whether the bounding box for the person in the task was drawn tightly around the person. If – for any bounding box in the image – any of the three annotators marked that the bounding box was not tight, the image was placed back in Stage 2 of the pipeline to be re-annotated.

A.4. Dataset Statistics and Breakdown

A.4.1 Attribute Representation

We detail the attribute breakdown for the remaining annotations in FACET. Table 9 details the statistics for the remaining person annotations. Table 10 shows the results of the robustness annotations with breakdowns on occlusion level and lighting condition.

A.4.2 Image statistics

We measure the statistics of images beyond specific attributes. Figure 9 shows the number of annotated people per image; less than one third of the images contain more than one person. Figure 10 shows the person box size as a fraction of total image size, broken down by the number of people in the image. All images in FACET are used for detection. Images with only one person are used for classification and visual grounding. **For masks, the 69k labeled masks span 18k people in 17k images of FACET. Each**

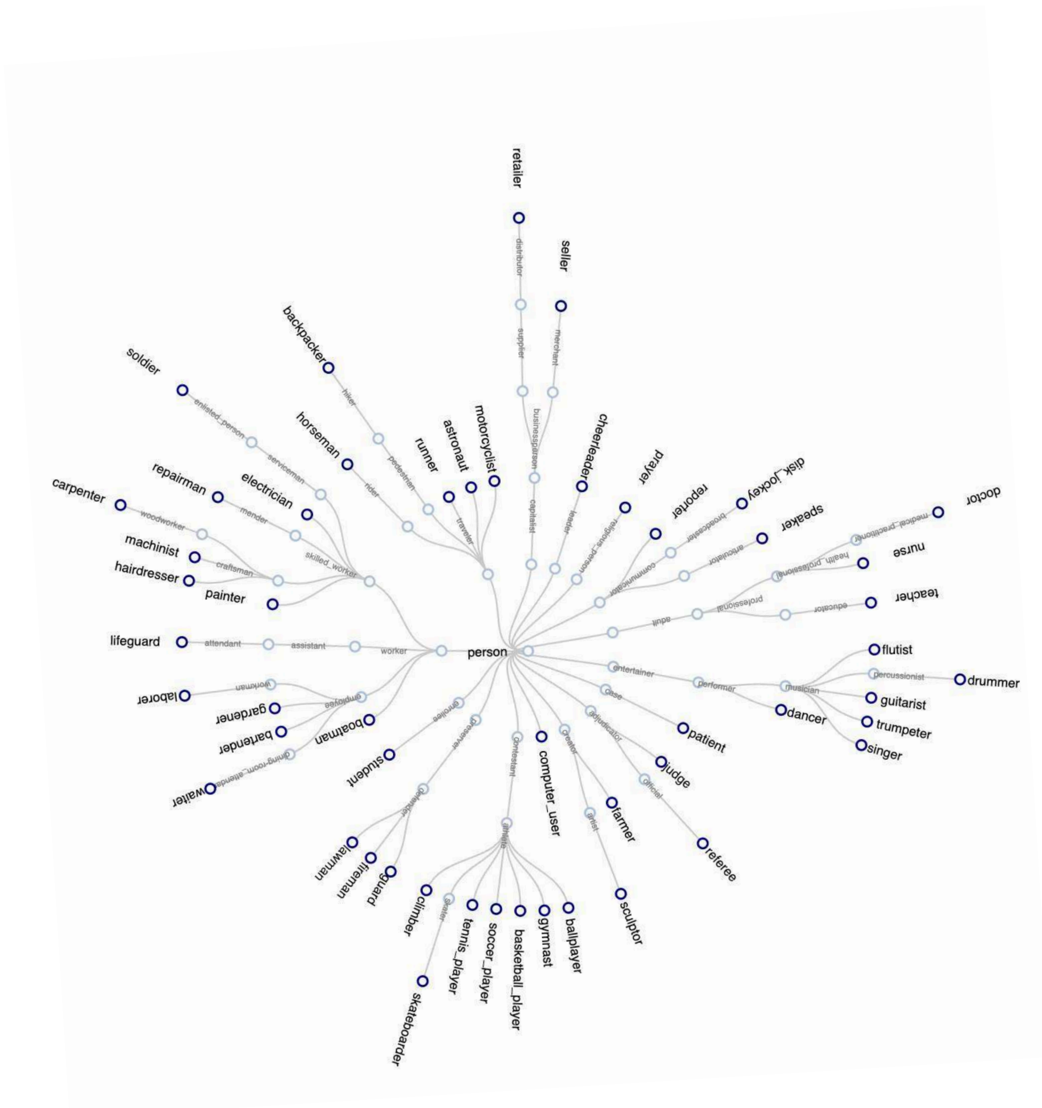


Figure 8: WordNet hierarchy of the FACET classes in relation to the `Person` synset. Classes are mapped to the `Person` synset (center) by their hyponyms (parents). Classes (leaves) are marked in blue. Grey nodes correspond to an intermediate hyponyms.

person with associated labeled masks has an average of 4 masks.

A.5. Evaluation

A.5.1 Dataset Setup

- For image classification, we limit the evaluation to examples in FACET that only contain one person. This helps alleviate ambiguities in performance. With this setup, we can consider the performance of the model on an image equivalent to performance of the model on the image for a specific set of attributes. There are 21k images in FACET that meet this criteria.
- For person and open world detection, we use all examples in FACET.
- For person segmentation, and the corresponding person detection baseline, we only use images and people inside each image that had a `person` mask - 11k people.
- For visual grounding, we only use examples in FACET with one person, as OFA predicts only one bounding box.

		people	%	images	%
Hair color	black	17k	34%	13k	42%
	blonde	3k	6%	3k	8%
	brown	11k	22%	9k	29%
	red/orange	547	1%	518	2%
	colored	269	1%	265	1%
	grey	2k	4%	2k	6%
	unknown	20k	40%	15k	46%
Hair type	wavy	9k	19%	8k	26%
	curly	761	2%	735	2%
	straight	19k	37%	15k	47%
	coily	458	1%	435	1%
	dreadlocks	296	1%	282	1%
	bald	1k	2%	965	3%
	unknown	23k	45%	16k	52%
Additional Annotations	eyeware	5k	11%	5k	15%
	headscarf	2k	5%	2k	6%
	tattoo	705	1%	672	2%
	cap	14k	29%	10k	33%
	facial-hair	6k	12%	5k	17%
	mask	3k	6%	2k	7%

Table 9: Statistics on the remaining person attributes: *hair color, hair type, presence of additional features* in FACET . Annotators could mark multiple hair colors and types for a single person.

	label	people	%	images	%
Lighting Condition	overexposed	941	2%	890	3%
	well-lit	40k	80%	27k	85%
	dimly-lit	11k	22%	9k	28%
	underexposed	1k	3%	1k	4%
	unknown	878	2%	849	3%
Visibility	minimal	7k	15%	7k	21%
	face	15k	30%	12k	38%
	torso	36k	73%	25k	78%

Table 10: Robustness annotations.

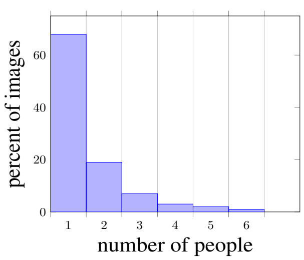
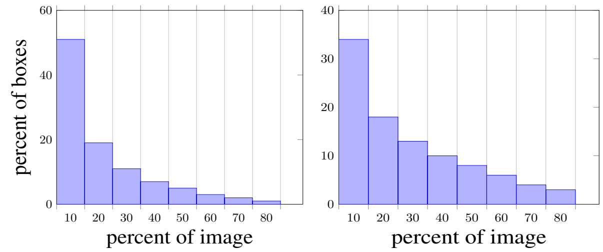


Figure 9: Histogram of number of people per image in FACET .



(a) all images (b) one person images

Figure 10: Histogram of person bounding box size as a percentage of total image size.

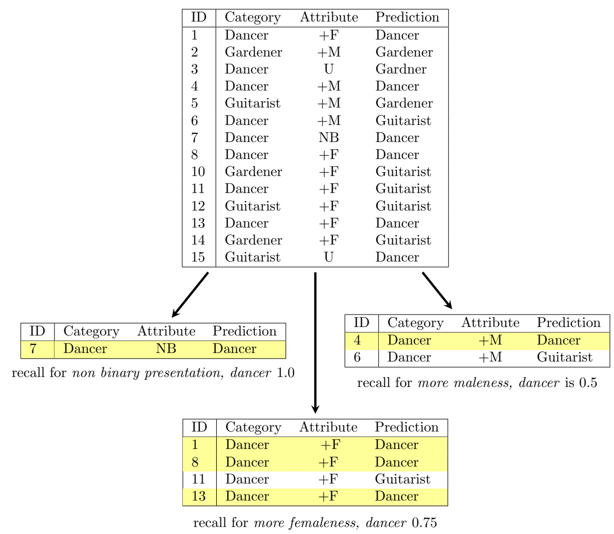


Figure 11: Example of how we score classification models for FACET .

A.5.2 Choice of Metric

We choose to focus on recall as it allows us to only consider examples with a specific demographic attribute or set of attributes. We choose to avoid a metric that would take into account false positives, as for some evaluations it is not clear what a false positive would mean. For example, for person detection, it is not obvious which demographic attribute a false positive would correspond to. *What demographic attributes would we consider a predicted false positive person to have?* While it might make sense for images with only one person to assume they had the same demographic attributes as the ground truth person in the photo, it is even less clear what the correct assumption would be to make if there were multiple people in the photo. To avoid this ambiguity, we focus on recall.

Classification We want to compare performance on a per-class basis, as overall performance metrics can hide disparities - i.e the model could have large biases but in opposite directions for two classes, which would yield a overall performance disparity of 0. We choose to look at each class separately. We don't want the metric to be influenced by the prevalence of the class for the group. We focus on the recall (at one) for the group and class for our evaluation. This is equivalent to the accuracy for the specific `(class, attribute)` pair. We note that we do not take into account true negatives of false positives. Figure 11 visualizes our metric. We note that there are multiple approaches to calculating a metric per class - e.g we could also look at the accuracy for the class when looking at all examples of the protected group, which is why detail the specifics of our considered metric.

Alignment with traditional fairness metrics The difference in recall we measure is equivalent to *equality of opportunity[41]* - larger differences in recall are further from equality of opportunity. *Equalized odds[41]* is an extension of this with analysis of true negative rate. The largest difference between M and F for CLIP is `retailers`, where M has a 3.8 higher TNR than F, suggesting that F are over-predicted as `retailers`. The largest difference between F and M for clip is `tennis player`, with a 3.0 higher TNR for F than M, suggesting that M are over-predicted as `tennis players`.

A.5.3 Classification

Experimental Setup In order to have maximum control over the experiment, we evaluate classification models on photos in FACET that only contain one annotated person. By filtering out images with > 1 person, we are left with 21k images. We look at the per class disparities between two groups only if both groups have at least 50 examples. We analyze CLIP based on recall.

ImageNet21k Pretraining As FACET categories overlap with ImageNet classes, we can evaluate ImageNet21k trained models out of the box. We take the max score over the FACET classes from the ImageNet class predictions. Table 12 shows a comparison of performance discrepancies across perceived age group for CLIP ViT B/32 and a ViT B/16 pre-trained on IN21k from [79].

Architecture Choice

A.5.4 Person Detection

We use a pre-trained FasterRCNN with a ResNet50 FPN backbone pretrained on COCO for person detection.

	Person Class	CLIP ViT B/32 #	Y	M	O	ViT B/16 IN21k #	Y	M	O
Top for CLIP	seller	1	**57.5**	72.8	**86.2**	9	**47.2**	53.4	**59.3**
	ballplayer	2	**60.6**	**75.5**	-	2	**57.6**	**77.4**	-
	guitarist	3	70.3	**80.2**	**65.5**	10	45.5	**47.9**	**36.4**
	speaker	4	**17.6**	28.5	**30.6**	4	**13.7**	25.7	**30.6**
	laborer	5	**49.0**	52.7	**61.7**	3	**48.1**	52.9	**66.0**
Top for ViT	painter	21	**56.5**	**51.0**	53.9	1	**37.0**	43.1	**57.8**
	ballplayer	2	**60.6**	**75.5**	-	2	**57.6**	**77.4**	-
	laborer	5	**49.0**	52.7	**61.7**	3	**48.1**	52.9	**66.0**
	speaker	4	17.6	28.5	**30.6**	4	13.7	25.7	**30.6**
	guard	7	**44.6**	32.9	-	5	**48.5**	31.7	-

Table 11: Per-class performance for CLIP and a ViT pre-trained on ImageNet 21k. A subset of FACET classes are shown. The perceived age groups with the highest performance discrepancy per class are bolded. (Y is *young*, M is *middle*, O is *older*). The top five classes with the biggest discrepancies per model are shown. # corresponds to the rank for class in terms of magnitude of the discrepancy. Lower number indicates larger discrepancy. We note that most of the classes are in the both of the model's top 10 classes with the largest discrepancies, and 2 classes are in both models top 5. Recall for class and perceived age group pairings with less than 50 samples are not reported.

	Person Class	ResNet IN21k #	Y	M	O	ViT IN21k #	Y	M	O
Top for ResNet	laborer	1	35.6	38.1	55.3	3	48.1	52.9	66.0
	guard	2	49.5	30.5		6	48.5	31.7	
	painter	3	38.9	35.9	53.9	1	37.0	43.1	57.8
	ballplayer	4	62.1	79.3		2	57.6	77.4	
	craftsman	5	67.2	78.4	81.8	12	74.6	78.7	81.8
Top for ViT	painter	3	38.9	35.9	53.9	1	**37.0**	43.1	**57.8**
	ballplayer	4	62.1	79.3		2	**57.6**	**77.4**	-
	laborer	1	35.6	38.1	55.3	3	**48.1**	52.9	**66.0**
	speaker	15	20.6	25.9	24.6	4	13.7	25.7	30.6
	guard	2	49.5	30.5		5	48.5	31.7	-

Table 12: Per-class performance for a ViT and ResNet pre-trained on ImageNet 21k. A subset of FACET classes are shown. The perceived age groups with the highest performance discrepancy per class are bolded. (Y is *young*, M is *middle*, O is *older*). The top five classes with the biggest discrepancies per model are shown. # corresponds to the rank for class in terms of magnitude of the discrepancy. Lower number indicates larger discrepancy. Recall for class and perceived age group pairings with less than 50 samples are not reported.

Additional Results Table 13 shows person detection results across perceived gender presentation and perceived age group.

Table 14 shows person detection results for a DETR[11] model with a ResNet50 backbone for perceived skin tone.

Demographic Group	mAR	$AR_{0.5}$	$AR_{0.75}$
perceived gender presentation			
– more stereotypically maleness	**74.4**	97.8	**83.1**
– more stereotypically femaleness	72.2	**97.9**	80.7
– outside of gender binary	71.2	**97.9**	76.8
perceived age group			
– younger	73.9	98.3	82.6
– middle	74.3	98.0	83.1
– older	**74.8**	**98.5**	**84.5**

Table 13: Average recall (AR) on FACET for a ResNet50 Faster R-CNN. Mean AR (mAR) averages across IoUs from 0.5 to 0.95 in increments of 0.05; $AR_{0.5}$ and $AR_{0.75}$ refer to IoU at 0.5 and 0.75.

Monk Skin Tone (MST)	mAR	$AR_{0.5}$	$AR_{0.75}$
1	**85.4**	**99.0**	**93.3**
2	84.6	98.8	92.1
3	84.4	98.7	91.6
4	84.2	98.6	91.3
5	84.0	98.6	91.2
6	84.0	98.7	91.2
7	83.8	98.6	91.1
8	84.1	98.6	91.5
9	83.6	98.6	90.9
10	82.8	98.2	90.1

Table 14: Average recall (AR) on FACET for a ResNet50-backbone DETR model. Mean AR (mAR) averages across IoUs from 0.5 to 0.95 in increments of 0.05; $AR_{0.5}$ and $AR_{0.75}$ refer to IoU at 0.5 and 0.75.

A.5.5 Person Segmentation

We use a MaskR-CNN[44] with a ResNet50 FPN backbone pretrained on COCO for person detection and instance segmentation. For this experiment, we only evaluate people in images if they have a mask annotated as `person` as well. This leaves us with 11k examples (people). For boxes, we compute the IoU of the predicted box to the human-labeled bounding box in FACET. For masks, we compute the IoU of the predicted mask to the Segment Anything-generated, annotator verified, mask in Segment Anything 1 Billion (SA-1B) [59]. Annotators verified and labelled the mask as `person`, and were instructed only to do so if the mask was around the entire person (like bounding boxes in FACET). Annotators did not make any updates to the mask boundary.

A.5.6 Open World Detection

Experimental Setup We use Detic [102] for open world detection. We use DETIC trained on IN21-k with a SWIN-B backbone. For the CLIP embeddings, we use the prompt '`a person who is a {}`' opposed to the '`a {}`' used in the original paper. As we focus on recall, we do not use a confidence threshold for DETIC's predictions. Similarly we allow multiple class predictions per box. We take the 100 top predictions per image to compute AR.

Additional Results Table 15 shows the per class disparities for all classes for perceived age group .

A.5.7 Visual Grounding

We evaluated OFA [93]. For OFA, we used the pretrained version OFA_{large} in the HuggingFace Transformers library [95]; we did not perform any additional finetuning. We used beam-search with 5 beams, *top-p*=0.6 and limited the generation to a maximum of 100 new tokens. We prompted OFA with the input (e.g. "Which region does the text `person class` describe?"). Because OFA produces a single bounding box per class per prompt, we only evaluated images that contained no more than one person instance per person class. 7858 images were excluded because they contained multiple instances per class. We show the average recall across different IoUs and for different perceived age group labels in Table 15.

	mAR			$AR_{0.5}$			$AR_{0.75}$			mAR			$AR_{0.5}$			$AR_{0.75}$		
	young	middle	older	young	middle	older	young	middle	older	young	middle	older	young	middle	older	young	middle	older
astronaut	64.0	**70.2**	-	80.0	**85.3**	-	60.0	**73.7**	-	0.0	**30.3**	-	0.0	**51.4**	-	0.0	**32.9**	-
backpacker	**45.4**	42.1	29.8	**55.3**	51.7	35.3	**47.4**	44.6	33.3	7.7	5.9	**11.5**	15.5	11.2	**22.0**	6.8	5.8	**9.8**
ballplayer	43.8	45.4	**61.8**	46.7	48.3	**63.6**	45.8	46.9	**63.6**	43.9	38.0	**58.0**	72.7	67.2	**100.0**	50.6	41.5	**80.0**
bartender	**81.7**	75.4	-	83.3	**85.4**	-	**83.3**	80.5	-	0.0	**12.5**	-	0.0	**25.0**	-	0.0	**8.3**	-
basketball player	61.2	**66.7**	-	69.0	**74.7**	-	64.9	**70.4**	-	26.2	**24.2**	-	**46.4**	41.3	-	25.0	**26.0**	-
boatman	**69.4**	59.9	64.0	**87.7**	79.1	82.3	**77.8**	65.3	68.8	1.4	**1.9**	1.5	4.5	6.2	**8.1**	0.6	**0.9**	0.0
carpenter	67.9	64.8	**81.7**	71.4	73.5	**91.7**	71.4	68.9	**87.5**	0.0	**2.4**	0.0	0.0	**5.6**	0.0	0.0	**2.2**	0.0
cheerleader	**13.3**	12.7	-	15.6	**13.5**	-	**14.8**	12.6	-	**20.0**	12.0	-	**41.9**	20.0	-	**16.1**	15.0	-
climber	**76.6**	74.4	67.5	**91.9**	92.4	75.0	**81.8**	81.0	75.0	0.0	**1.2**	0.0	0.0	**2.5**	0.0	0.0	**0.8**	0.0
computer user	**72.9**	66.3	68.4	**81.0**	77.8	73.7	**76.3**	67.5	68.4	7.9	5.9	**8.6**	12.9	10.3	**14.3**	**8.2**	6.2	7.1
craftsman	44.5	47.0	**56.9**	48.6	52.1	**61.9**	44.8	48.3	**58.6**	33.8	39.1	**40.7**	55.2	62.6	**66.9**	37.9	45.7	**47.6**
dancer	**77.2**	71.1	75.6	**91.4**	85.4	87.5	**83.7**	77.5	78.1	**37.6**	32.0	24.3	**68.8**	55.7	57.1	**39.0**	37.1	28.6
disk jockey	**77.2**	68.4	-	**82.1**	78.8	-	**79.1**	72.5	-	3.5	**3.8**	-	**6.8**	6.5	-	**4.1**	3.0	-
doctor	74.6	**77.4**	75.7	86.2	**88.7**	81.0	76.6	**79.8**	78.6	33.6	30.9	**38.0**	55.2	52.1	**60.8**	40.2	33.6	**45.1**
drummer	19.9	26.3	**34.2**	24.9	34.7	**41.8**	19.7	27.6	**35.8**	**5.0**	3.8	0.7	**9.2**	8.1	1.7	**4.6**	3.3	0.0
electrician	**56.3**	51.4	48.6	**62.8**	62.5	57.1	**62.8**	54.4	57.1	0.0	**1.2**	0.0	0.0	**1.6**	0.0	0.0	**1.6**	0.0
farmer	81.5	81.1	**85.4**	95.9	96.6	**99.1**	86.9	88.4	**93.0**	6.2	5.0	**6.6**	12.8	9.9	**13.7**	5.1	4.4	**5.5**
fireman	**86.3**	76.4	76.4	**96.2**	90.1	85.7	**90.4**	82.6	85.7	14.0	14.7	**22.0**	26.7	**32.9**	60.0	13.3	12.5	**20.0**
flutist	32.1	40.5	**51.0**	35.4	47.5	**54.8**	35.4	43.7	**54.8**	**15.0**	10.5	11.7	**31.8**	19.9	20.8	9.1	9.9	**12.5**
gardener	82.3	78.6	**86.8**	98.3	94.7	**100.0**	90.0	84.4	**97.3**	11.9	18.3	**27.9**	32.6	40.1	**58.1**	7.0	14.6	**24.2**
guard	81.9	80.2	**88.5**	94.3	90.6	**97.5**	89.4	87.2	**95.0**	14.1	15.2	**19.2**	34.0	31.9	**38.5**	9.6	**12.5**	11.5
guitarist	75.9	79.3	**79.5**	90.5	93.7	**95.1**	80.0	84.4	**86.3**	19.8	19.6	**32.0**	38.7	35.5	**56.7**	18.5	20.6	**36.7**
gymnast	**87.7**	85.5	-	**96.2**	95.6	-	**92.4**	89.9	-	**10.0**	8.5	-	**19.1**	17.1	-	**9.8**	9.1	-
hairdresser	76.8	**79.4**	79.0	94.1	**96.9**	92.9	82.4	79.9	**83.3**	**15.2**	13.3	12.1	**28.0**	24.3	23.5	12.0	12.8	**14.7**
horseman	**70.9**	62.1	64.5	**85.4**	75.7	80.0	**77.2**	67.9	70.0	13.4	**14.5**	11.0	36.5	**38.8**	30.0	3.8	**5.4**	0.0
judge	25.7	**31.3**	28.3	28.6	**35.3**	33.3	28.6	**33.8**	33.3	–	**10.4**	0.0	–	**25.0**	0.0	–	**3.6**	0.0
laborer	**75.3**	73.1	74.4	88.4	85.8	**86.1**	**79.9**	78.9	**79.9**	23.1	21.9	**28.9**	44.0	46.2	**58.6**	22.0	17.4	**24.3**
lawman	**71.5**	70.1	67.1	**79.0**	77.7	74.3	**75.5**	74.6	70.6	20.2	21.1	**22.6**	42.0	43.1	**46.2**	18.3	18.8	**21.5**
lifeguard	41.8	46.1	**52.5**	51.7	54.9	**62.5**	47.5	49.8	**62.5**	**7.5**	7.0	0.0	**19.7**	17.9	0.0	2.8	**5.2**	0.0
machinist	**60.0**	49.9	41.1	**63.9**	56.5	44.4	**63.9**	52.2	44.4	21.7	21.3	**23.3**	34.8	35.5	**41.7**	**26.1**	25.0	25.0
motorcyclist	**57.9**	52.7	51.9	**81.6**	78.2	69.2	**60.9**	54.2	57.7	**21.9**	15.5	19.2	**50.0**	37.0	37.5	12.3	9.6	**20.8**
nurse	**83.4**	81.5	81.7	**95.6**	93.9	91.3	**90.5**	86.1	82.6	31.8	24.8	**34.5**	**52.2**	43.6	50.0	37.2	26.0	**40.0**
painter	54.0	58.9	**68.6**	60.8	66.3	**73.8**	58.2	62.3	**73.8**	**18.0**	15.6	17.7	**30.1**	29.3	27.6	**23.3**	16.1	20.4
patient	64.1	66.9	**67.1**	**87.0**	85.6	86.5	65.6	**69.2**	68.3	**28.5**	26.5	26.6	**50.3**	47.6	45.2	**29.7**	27.6	28.0
prayer	82.8	83.0	**85.2**	**96.0**	95.2	95.2	89.0	**89.5**	**89.5**	0.0	2.7	**2.8**	0.0	**5.5**	4.3	0.0	1.8	**2.9**
referee	70.2	77.5	**84.9**	75.5	85.3	**91.4**	73.6	80.9	**88.6**	19.6	20.4	**21.4**	40.8	40.1	**45.7**	16.3	19.8	**22.9**
repairman	**71.2**	61.7	65.2	**77.6**	69.7	71.0	**75.0**	65.5	69.6	**20.1**	17.9	17.0	**39.5**	32.8	30.4	18.4	**19.1**	17.4
reporter	21.7	22.9	**25.0**	23.7	25.7	**29.2**	22.4	23.7	**25.0**	**9.2**	5.2	4.5	**19.7**	13.0	6.9	**7.0**	3.9	3.4
retailer	33.3	35.0	**52.2**	40.9	43.2	**59.5**	33.6	38.6	**54.1**	1.0	2.5	**3.1**	2.8	6.5	**6.9**	0.0	1.5	**3.4**
runner	90.9	85.9	**91.1**	99.2	95.2	**100.0**	97.7	90.6	**100.0**	7.5	**8.3**	0.0	**21.6**	21.1	0.0	3.9	**4.7**	0.0
sculptor	74.5	73.0	**85.0**	81.8	82.4	**95.8**	77.3	77.6	**83.3**	**2.4**	2.3	0.0	**5.9**	5.1	0.0	0.0	**3.1**	0.0
seller	73.0	73.2	**74.5**	87.4	87.0	**88.5**	**82.0**	79.0	80.8	7.8	8.2	**9.8**	16.8	16.6	**21.1**	6.2	6.9	**8.8**
singer	80.6	80.9	**85.0**	88.8	88.2	**96.1**	85.0	85.5	**88.2**	**5.1**	3.9	1.7	**10.3**	7.3	5.8	**4.5**	3.6	0.0
skateboarder	40.7	**43.1**	-	45.1	**46.5**	-	43.4	**46.1**	-	21.5	**23.9**	–	46.9	**49.7**	–	18.8	**22.1**	–
soccer player	81.8	**82.9**	-	90.6	**91.7**	-	86.2	**87.1**	-	**26.7**	22.6	–	**49.1**	42.8	–	**24.6**	21.4	–
soldier	**65.1**	63.4	51.4	**72.8**	72.7	56.8	**69.7**	68.3	54.1	**16.3**	**16.3**	1.2	**40.0**	33.8	6.2	9.2	**14.5**	0.0
speaker	83.0	80.8	**85.1**	89.3	88.5	**93.1**	87.7	85.6	**89.1**	2.0	1.7	**2.1**	**4.6**	3.2	3.7	**1.9**	1.7	1.6
student	60.6	**71.1**	-	69.8	**80.9**	-	64.6	**74.4**	-	**29.0**	25.3	0.0	**51.9**	44.7	0.0	**33.8**	25.0	0.0
teacher	**83.4**	81.0	80.0	**96.6**	90.3	87.5	**93.1**	85.8	87.5	**28.1**	22.2	15.0	**51.6**	39.8	50.0	**29.0**	24.8	0.0
tennis player	**94.2**	93.8	-	**98.9**	**98.9**	-	97.2	**97.8**	-	32.5	**33.8**	-	60.0	**62.2**	-	32.7	**34.4**	-
trumpeter	22.8	29.5	**38.4**	26.7	34.8	**45.5**	25.6	31.4	**38.2**	**5.3**	5.1	3.6	**11.6**	10.3	5.1	2.3	**5.7**	5.1
waiter	76.2	**77.6**	-	92.4	**92.9**	-	**83.3**	82.5	-	**5.2**	4.2	-	**10.4**	8.6	-	**4.2**	4.0	-
avg	64.6	64.0	**68.2**	74.1	74.4	**76.4**	68.6	67.9	**72.4**	**14.7**	14.5	14.0	**28.5**	27.8	26.2	13.8	14.1	**15.0**

(a) Results for Detic (b) Results for OFA

Table 15: The average recall (AR) results for Detic (detection) and OFA (visual grounding) across the 52 person-related classes for each perceived age group label. The highest recall numbers are bolded.

B. Data Card

We provide a data card for FACET, following the guidance of [48].

FACET

`https://facet.metademolab.com`

FACET is a large, publicly available evaluation set of 31,702 images for the most common vision problems - **image classification**, **object detection**, **segmentation**. People in FACET are annotated with person-related attributes such as **perceived skin tone** and **hairtype**, **bounding boxes** and labeled with fine-grained **person-related classes** such as *disk jockey* or *guitarist*.

Overview	
Publisher	Meta AI Research, FAIR
Authors	Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, Candace Ross
Contact	facet@meta.com
Funding & Funding Type	Industry
License	Custom license, see dataset download agreement
Applications	
Dataset Purpose	Evaluate computer vision models to detect potential fairness concerns
Key Application	*Computer Vision, Fairness and Robustness*
Primary Motivations	Give researchers a tool to help understand model fairness. Allow researchers to investigate how the demographic attributes of a person in the photo correlates with model performance. FACET supports common vision tasks, with annotations for classification detection, and segmentation.
Intended Audience	Researchers aiming to detect potential fairness concerns and biases in their trained vision models.
Suitable Use Case	**FACET is for evaluation only.**
Data Type	
Primary Data Type	Images
Primary Annotation Types	Manually gathered annotations for: • Bounding boxes • Category labels for the bounding boxes • A series of demographic, robustness, and additional attributes for the person in the bounding box. • Manually annotated labels for mask from Segment Anything 1 Billion (SA-1B) [59]. This masks were automatically generated by the Segment Anything Model (SAM).

FACET Data Card	
Data SnapShot	• **31,702** images • exhaustive annotations for **49,551** people • **52** categories for people that include occupations, athletes, artists, etc • **13** attributes annotated for person including demographic attributes such as *perceived gender presentation* robustness annotations such as *lighting condition* and additional attributes such as *hair color* • **3** mask labels *person, clothing, hair* for masks. Masks and mask labels are not exhaustive. 17k people in 14k images have labelled masks. Additional unlabeled masks from SA-1B are compatible with FACET .
Data Sources	Images come from SA-1B[59] .

FACET Data Card	
Annotation format	JSON files of COCO formatted annotations for the bounding boxes and masks are provided. A CSV containing the annotations per person is be provided. Each item in the annotation file contains: 1. Reference information: • `filename` • `person_id`: unique integer representing the annotation 2. Task information: • `class1`: This is the primary category the person matches. *Cannot be None.* • `class2`: This is the secondary category the person matches. *Can be None.* • `bounding_box`: Person bounding box. • `masks`: Each item will contain the `category` and `mask`. `Category` will be one of `person, hair, clothing`. There are not masks for every person/image. 3. Demographic Attribute annotations. • *perceived gender presentation* : All of the following annotations will given in a binary fashion: `[with_more_femaleness, with_more_maleness, nonbinary_presentation, gender_presentation_unknown]` • *perceived skin tone* : Each annotators annotations are considered per MST in a binary fashion. Annotations from all annotators are summed into a single value per MST, so the value at MST_i may be greater than 1. Values will be given for all of the following: [MST_1, ..., MST_{10} apparent_skin_tone_unknown] • *perceived age group* : all of the following annotations are included in a binary fashion: `[young, middle, older, age_presentation_unknown]` 4. Additional Attribute information: All binary values. • *hair color:* `[black_hair, red_hair, blonde_hair, brown_hair, colored_hair, grey_hair, hair_color_unknown]` • *hair type:* `[wavy, curly, coily, straight_hair, bald, dreadlocks, hair_type_unknown]` • *other items:* `[eyewear, headscarf, tattoo, cap, facial_hair, mask]` 5. Robustness Annotations: All binary values. • *lighting condition:* `[lighting_unknown, overexposed, underexposed, well_lit, dimly_lit]` • *visibility:* `[minimal_visible, torso_visible, face_visible]`

C. FACET CrowdWorkSheets

To further describe our annotation process, we answer the questions posed in CrowdWorkSheets[21].

C.1. Task Formulation

At a high level, what are the subjective aspects of your task? Annotating the *perceived* attributes of a person is by nature subjective. For perceived skin tone we expected the annotations would be subjective and have high variance. To account for this, we gather annotations from three annotators and release the cumulative results of all three. For subjectivity across the other attributes and labeling classes, we provided annotators with diverse representations of each attribute or class in the guidelines to try to minimize annotator bias.

What assumptions do you make about annotators? How did you choose the specific wording of your task instructions? What steps, if any, were taken to verify the clarity of task instructions and wording for annotators? To qualify for the annotation task, annotators had to pass a strong English requirement. For the annotation of perceived skin tone only, we had a more lenient English requirement to increase the diversity of the annotators, and additionally translated the annotation instructions into Spanish.

As we were annotating images, we provided visual examples for all of the annotations and classes. We sourced multiple examples per attribute (e.g brown hair) and class (e.g doctor), with at least one example for someone with more stereotypical maleness with the attribute and someone with more stereotypical femaleness with the attribute. For classes, we sourced multiple examples of someone who would qualify for a given class (*e.g for dancer we sourced images of both a ballerina and a break-dancer*). For given examples for the Monk Skin Tone scale, we sourced four examples per MST value, and attempted to capture some of the diversity within a specified MST value.

What are the precise instructions that were provided to annotators? The goal of the project is to build a dataset that helps determine if Computer Vision models have biases based on the apparent attributes of the person in the photo. We are creating an image classification dataset that also contains labels of the apparent protected attributes of the people in the image. The dataset is for evaluation only, and is to help better analyze and detect potential biases. The protected attributes will not in any way be used for training a model. We are not collecting any biometric information about the people in the photos.

1. **Target category classification:** Given an image, and a target category, we aim to determine if the image is a good representation for the category. The annotators will mark whether or not there is a person in the photo matching the category, and if so if there are ≤ 5 people who match this category. The categories will be all people related - such as doctor, soccer player, etc. Multiple images will be shown per task to annotate. The default response will be 'No person matches this category'.

2. **Bounding boxes and classification labels for people:** Given an image, draw bounding boxes around all people who match any of the list of categories. For each bounding box around a person, mark which category they belong to. If they belong to multiple categories, you should mark the second category under 'secondary category'.

3. **3. Apparent skin tone annotations** Given an image, with a bounding box around a person, annotate the person's apparent skin tone. You may select as many skin tones from the list as you feel appropriate. If it is not possible to tell the skin tone from the photo, please mark cannot be determined. Please select at least two values for the skin tone, and make sure that the values that you select are consecutive. If it is too hard to determine the annotation, mark the values it appears and cannot be determined. Zoom in (option + mouse scroll) as necessary in order to determine the skin tone.

4. **4. Apparent attribute annotations** Given an image, with a bounding box around a person, annotate the given apparent attributes of the person. For each category, see the examples given. If it is not possible to determine the attribute from the photo, please mark cannot be determined. Apparent lighting condition is on the person: Please indicate how the lighting is with respect to the person in the bounding box. If the lighting is between two categories, mark both.

C.2. Selecting Annotations

Are there certain perspectives that should be privileged? If so, how did you seek these perspectives out? No. N/A

Are there certain perspectives that would be harmful to include? If so, how did you screen these perspectives out? Harmful perspectives would include annotators who had a clear bias in their annotations. We screened these perspectives out by using training, and only including production raters who had high accuracy on the training set. Annotators with consistent bias would likely not have been able to get a high enough accuracy on the training to graduate.

Were sociodemographic characteristics used to select annotators for your task? If so, please detail the process. If you have any aggregated sociodemographic statistics about your annotator pool, please describe. Do you have reason to believe that sociodemographic characteristics of annotators may have impacted how they annotated the data? Why or why not? We sourced geographically diverse annotators from the following 7 countries during our annotation process: United States, Philippines, Egypt, Colombia, Taiwan, Spain and Kenya. The breakdown of annotators per region is shown in Figure 4 in the main text.

If you have any aggregated socio-demographic statistics about your annotator pool, please describe. Do you have reason to believe that sociodemographic characteristics of annotators may have impacted how they annotated the data? Why or why not? Other socio-demographic statistics about our annotator pool were not available.

Consider the intended context of use of the dataset and the individuals and communities that may be impacted by a model trained on this dataset. Are these communities represented in your annotator pool? The FACET benchmark is to be used for evaluation purposes only. The underlying images in FACET are geographically diverse. To incorporate geographic diversity into our annotation process, we sourced annotators from 7 countries across regions.

C.3. Platform and Infrastructure Choices

What annotation platform did you utilize? At a high level, what considerations informed your decision to choose this platform? Did the chosen platform sufficiently meet the requirements you outlined for annotator pools? Are any aspects not covered? We used a proprietary annotation platform.

What, if any, communication channels did your chosen platform offer to facilitate communication with annotators? How did this channel of communication influence the annotation process and/or resulting annotations? For Stage 2 (drawing and labeling bounding boxes for person classes), labelers' annotations were compared to a golden set and were required to achieve IoU above 85% to pass. After these training stages, annotations were manually reviewed and the annotators were given feedback for improvement. Following this, if annotators had high quality labels when spot-checked, they graduated to annotating images for the final benchmark.

We provided annotators individualized feedback during their training for drawing bounding boxes on a daily basis. Our vendor communicated to annotators common types of mistakes that we witnessed during training, and the corresponding corrections.

We provided annotators individualized feedback during their training for drawing bounding boxes. Our vendor communicated to annotators common types of mistakes that we witnessed during training, and the corresponding corrections.

How much were annotators compensated? Did you consider any particular pay standards, when determining their compensation? If so, please describe. Annotators were compensated with an hour wage set per country.

C.4. Dataset Analysis and Evaluation

How do you define the quality of annotations in your context, and how did you assess the quality in the dataset you constructed? For each task, annotators were first placed into training for the task. They were asked to annotate a large number of examples per task. We hand annotated the same examples, and using our annotations as the ground truth measured the accuracy per annotator. Annotators were graduated from training when their accuracy reached above a given threshold. For the task requiring annotators to draw bounding boxes around people, annotators were only graduated after we manually spot checked the annotator's bonding boxes to ensure quality. During the perceived skin tone annotation task, we asked annotators if they agreed with the class label, and grade the quality of the given bounding box. If one of the three annotators disagreed with the class label or bounding box, the annotation was removed, and the image added to the queue of images for task 2 (drawing bounding boxes).

Have you conducted any analysis on disagreement patterns? If so, what analyses did you use and what were the major findings? We pointed out common mistakes during weekly meetings with the vendor. While in training, we noticed consistent mistakes among annotators that we corrected before graduation. The most common mistake was around drawing the bounding boxes: many annotators during training would draw bounding boxes that included objects the person was holding *e.g guitar*. With the weekly meetings and individualized feedback, we were able to address this.

How do the individual annotator responses relate to the final labels released in the dataset? For perceived skin tone only, we sourced three annotations per person in the dataset. We release the annotations from all three annotators, giving a distribution over perceived skin tone per person in the dataset. We believe that a distribution more accurately describes a person's perceived skin tone than a single value.

C.5. Dataset Release and Maintenance

Do you have reason to believe the annotations in this dataset may change over time? Do you plan to update your dataset? At this time we do not plan to have updates for this dataset. We will allow users to flag any images that may be objectionable content, and remove objectionable content if found.

Are there any conditions or definitions that, if changed, could impact the utility of your dataset? The FACET benchmark contains examples for many different types of professions, athletes, artists, etc. If over time the way these occupations look shifts, this could impact the dataset. As a concrete example, there are a number of images in the dataset that were taken since the beginning of the COVID-19 pandemic. Many doctors and nurses in the dataset are wearing much more PPE than in images of doctors and nurses from before the COVID-19 pandemic.

Will you attempt to track, impose limitations on, or otherwise influence how your dataset is used? If so, how? The FACET benchmark is for evaluation purposes ONLY. Using FACET annotations for training is strictly prohibited. Users must agree to the terms of use before downloading the dataset.

Were annotators informed about how the data is externalized? If changes to the dataset are made, will they be informed? No. No.

Is there a process by which annotators can later choose to withdraw their data from the dataset? If so, please detail. No.

D. Fine-grained dataset statistics

FACET Class Statistics																	
		Perceived Gender Presentation			Perceived Skin Tone										Perceived Age Group		
Person Class	Total	stereotypical maleness	stereotypical femaleness	non-binary presentation	1	2	3	4	5	6	7	8	9	10	younger	middle	older
lawman	4609	3768	403	3	560	2363	2881	2642	1825	1215	615	322	166	74	387	3151	144
laborer	3030	2208	378	1	112	577	887	1171	1171	1269	844	508	291	136	297	1643	193
boatman	2147	1074	742	5	137	741	991	995	794	573	296	134	86	31	482	906	147
guard	1851	1597	121	4	306	1045	1208	1047	714	470	246	119	62	27	317	1181	48
backpacker	1738	1006	458	4	167	771	1010	951	761	475	193	93	50	28	361	842	53
basketball player	1680	1479	134	2	307	869	991	845	632	461	413	359	332	165	492	1056	3
tennis player	1663	1058	488	0	147	805	1152	1262	1002	617	234	126	90	57	360	1081	6
farmer	1632	823	539	1	50	208	335	466	635	816	681	450	216	87	129	844	227
soldier	1561	1336	75	0	204	766	892	802	578	463	281	130	66	22	237	972	39
singer	1518	1013	428	14	240	824	1013	931	677	399	184	140	93	46	357	984	89
dancer	1475	510	812	10	207	644	863	798	716	431	214	113	67	26	567	644	32
speaker	1470	1119	282	1	152	789	1093	1050	755	392	182	92	49	25	134	987	207
motorcyclist	1468	822	302	4	55	384	518	583	539	478	249	104	59	22	213	577	56
repairman	1430	1187	54	1	65	480	681	815	753	676	373	150	60	18	126	836	78
seller	1342	699	533	6	74	359	555	705	757	642	379	192	74	28	205	758	184
ballplayer	1316	1145	62	1	104	584	743	781	635	473	253	168	112	51	214	834	12
guitarist	1279	1115	87	3	138	678	843	816	596	330	139	73	50	26	233	802	116
computer user	1267	597	322	2	176	641	818	785	608	358	152	71	34	12	258	449	24
soccer player	1233	1102	34	1	113	521	692	711	559	364	200	126	128	76	322	732	5
craftsman	1127	785	220	4	75	321	467	598	631	627	389	210	92	36	117	599	188
nurse	1124	322	535	3	115	368	505	536	529	399	188	86	27	6	169	547	24
drummer	1006	744	162	3	114	428	534	483	388	331	222	150	99	42	256	530	68
skateboarder	1000	818	88	1	82	468	635	650	463	281	136	62	37	14	360	465	1
painter	983	590	251	0	77	318	460	530	506	420	246	123	56	22	168	435	129
fireman	933	674	34	0	68	270	358	391	237	192	77	22	10	7	55	512	14
patient	896	408	275	0	75	280	389	472	486	444	242	102	41	16	131	368	127
horseman	884	491	290	1	152	538	592	484	287	127	54	26	12	5	181	512	22
doctor	861	361	313	1	86	343	450	462	410	284	145	69	21	6	107	441	43
prayer	810	444	265	3	58	223	307	355	394	357	195	99	51	18	104	358	124
referee	776	694	38	1	88	417	539	547	374	186	80	38	23	10	54	584	35
student	747	379	247	1	92	241	322	365	367	316	163	93	62	26	319	264	5
runner	654	469	117	3	88	320	415	379	291	126	45	23	27	16	134	403	19
gymnast	635	252	316	1	116	348	424	366	322	145	43	28	17	6	300	233	2
retailer	561	296	234	0	53	198	301	298	298	196	90	41	14	5	114	332	39
climber	551	355	92	2	59	231	306	301	251	155	73	26	13	8	107	261	4
trumpeter	530	451	36	3	63	308	336	304	212	145	74	41	37	18	89	316	56
lifeguard	529	398	62	0	20	160	232	286	229	186	103	52	28	8	118	273	8
electrician	505	415	7	0	9	100	140	188	182	175	101	52	35	9	47	270	9
gardener	499	266	173	1	45	187	257	265	235	197	108	56	33	16	66	245	79
reporter	473	302	145	1	75	281	324	269	204	116	50	18	13	4	77	302	24
hairdresser	461	342	85	3	32	143	209	257	242	237	145	75	35	17	69	294	43
machinist	413	329	30	0	33	173	223	252	191	168	89	34	20	7	42	241	20
cheerleader	410	78	314	0	77	191	292	268	205	88	38	18	12	3	246	117	5
waiter	350	204	109	1	34	184	245	220	177	120	51	24	18	7	68	224	7
disk jockey	318	228	27	1	43	162	200	194	127	77	37	27	20	10	67	167	2
flutist	312	247	41	0	38	152	192	184	154	118	77	43	16	4	50	189	32
astronaut	289	165	14	0	15	72	89	78	58	18	2	0	0	2	5	158	2
carpenter	268	230	7	0	11	82	124	147	129	131	87	52	25	9	20	160	27
sculptor	240	187	21	0	10	76	104	120	107	107	78	50	24	5	24	144	27
teacher	216	116	76	1	28	104	141	142	108	76	36	16	10	4	31	150	9
judge	101	67	28	0	11	50	76	71	44	21	6	3	1	0	8	71	12
bartender	57	37	14	0	5	27	42	36	29	19	7	3	1	1	7	41	1

Table 17: Number of people for each person class and demographic group in FACET.

FACET Mask Statistics			
	person	clothing	hair
perceived gender presentation			
with stereotypical maleness	6608	32103	3788
with stereotypical femaleness	4127	18136	3346
non-binary presentation	50	223	36
cannot be determined	72	193	13
perceived skin tone			
MST 1	2198	10687	1389
MST 2	5154	24328	3496
MST 3	6121	28825	4263
MST 4	5651	26583	3889
MST 5	4849	22738	3349
MST 6	3816	17931	2452
MST 7	2542	11845	1544
MST 8	1619	7564	922
MST 9	1216	5727	666
MST 10	521	2481	293
cannot be determined	2839	11844	1611
perceived age group			
younger	4145	19440	3107
middle	5443	25458	3319
older	1134	5352	733
cannot be determined	135	405	24
Hair color			
black	4053	18137	3323
brown	2726	12205	2267
blonde	1024	4633	952
red/orange	148	674	136
colored	84	340	96
grey	747	3519	559
cannot be determined	2885	14863	485
Hair type			
wavy	2090	9526	1897
curly	241	1141	253
straight	5141	22109	4395
coily	178	750	158
dreadlocks	113	522	109
bald	265	1167	81
Unknown	3626	19129	905
Additional attribute			
eyeware	1509	6993	957
headscarf	665	3634	256
tattoo	184	926	143
cap	3305	18209	797
facial hair	1511	7382	963
mask	591	3271	377

Table 18: Number of masks per type per demographic and additional attributes in FACET. For perceived skin tone, hair color, hair type, and additional attributes a person in FACET can be marked with multiple values, so the sum of the masks over the group of attributes greater than the total number of masks.