Supplementary Materials

A. Max-min clustering algorithm for seeds filtering

We use an algorithm based on max-min clustering as described in Alg. 1 to filter the seeds. Refer to Sec. 3.1 for more details.

Algorithm 1 Max-min clustering seeds filtering

Require: Base seeds S_B, filtered seeds S_F, face mesh model M, seeds requied n, DemographicGroup(DG) = {"WM", "WF", "BM", "BF", "AM", "AF"}

$S_F \leftarrow \{S_B[0]\}$
while $i \leq n$ **do**
  for $s \in S_B \setminus S_F$ **do**
    $D(s) = \min_{\forall f \in S_F, \forall g \in DG} = ||M(I_s^g) - M(I_f^g)||_2$
  end for
  S_F.add($\arg\max_s D(s)$)
  $i = i + 1$
end while

B. Image annotation interface for identity comparison

We collect human annotations using *Amazon SageMaker Ground Truth* and we show an example of interface in Fig 1. Refer to Sec. 3.5 for details.

C. Example image before & after face segmentation and background removal

We show examples of images before and after segmentation and background removal in Fig. 2. Refer to Sec. 3.3 for details.

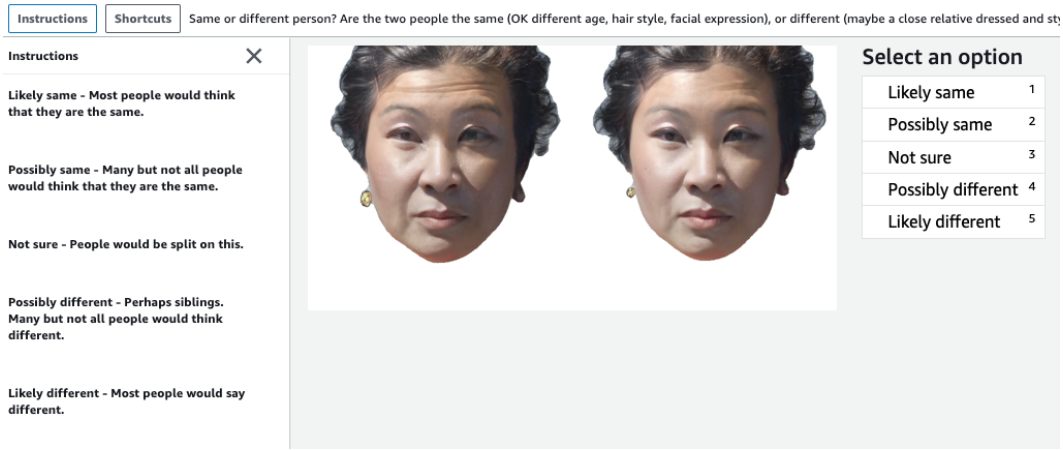

Figure 1. **Human annotation interface.** We give annotators a pair of face images and ask them to choose from one of the following options: {*'likely same', 'possibly same', 'not sure', 'possibly different' and 'likely different'*}.

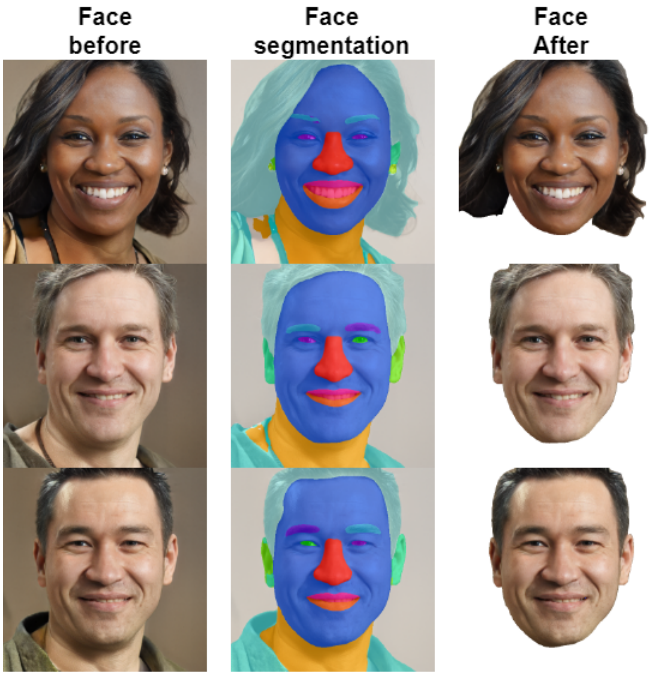

Figure 2. **Face segmentation & background removal**. To reduce noise in both model and human's prediction, we perform face segmentation and background removal to all the images.

D. Single image annotation results for skin tone and uncanny.

We asked annotators to label image realism (“uncanniness ” in our surveys) as well as race (“skin color” in our surveys), the results are in Fig 3. This allows us to throw away any images that are unrealistic or have significant artifacts. Refer to Sec. 4.1 for details.

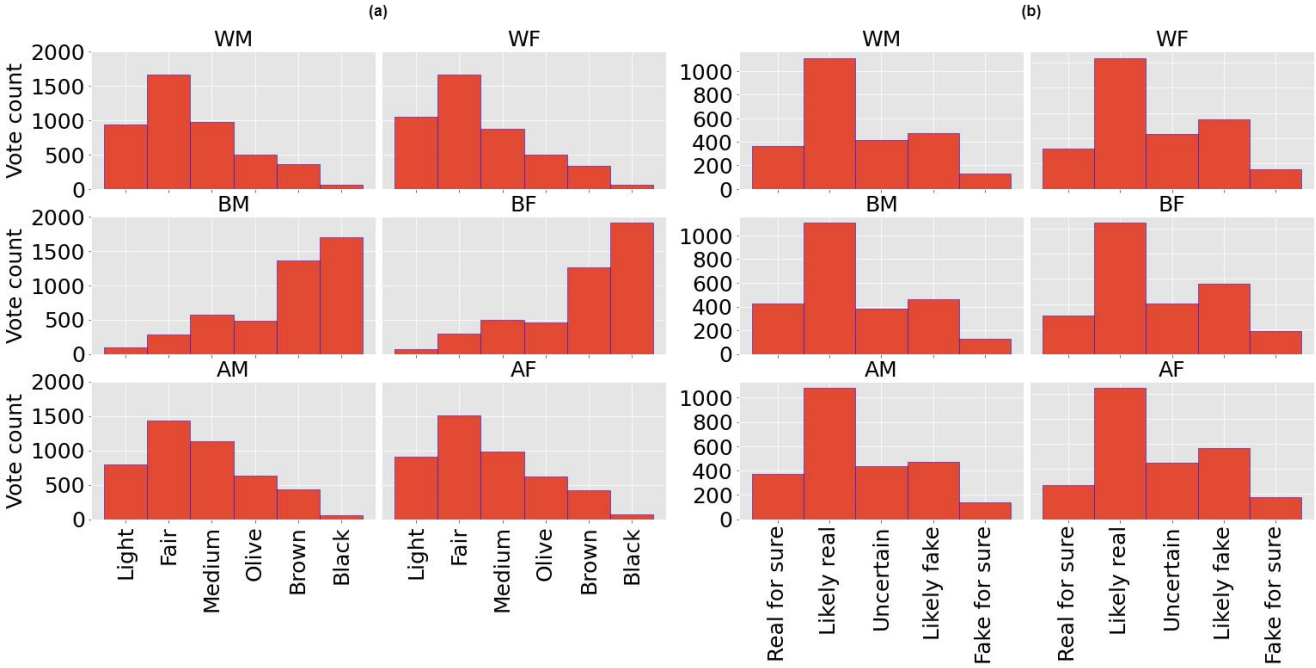

Figure 3. **Single image annotation**. **(a)**: Results for skin color, where annotators give scores per image ranging from ”light” to ”black”. **(b)**: Results for uncanniness, where annotators give score per image ranging from ”real for sure” to ”fake for sure”. We split results by demographic groups.

E. Identity similarity scores for other attributes and models

We feed face image pairs to three different pretrained popular public face recognition models: a ResNet34 trained on MS1MV3 using ArcFace, a ResNet34 trained on Glint360k using ArcFace, and a SFNet20 trained on VGGFace2 using SphereFace. Note that since we don't have ground truth labels for "age" and "expression", we instead use the results from single image annotation(see Sec. 4.1) to assign them age/expression group: group $\{0, 1, 2, 3, 4\}$ represent images whose scores are in $\{[0, 0.8), [0.8, 1.6), [1.6, 2.4), [2.4, 3.2), [3.2, 4]\}$ respectively. The results are shown in Fig 4, 5, 6. Refer to Sec. 4.1.1 for details.

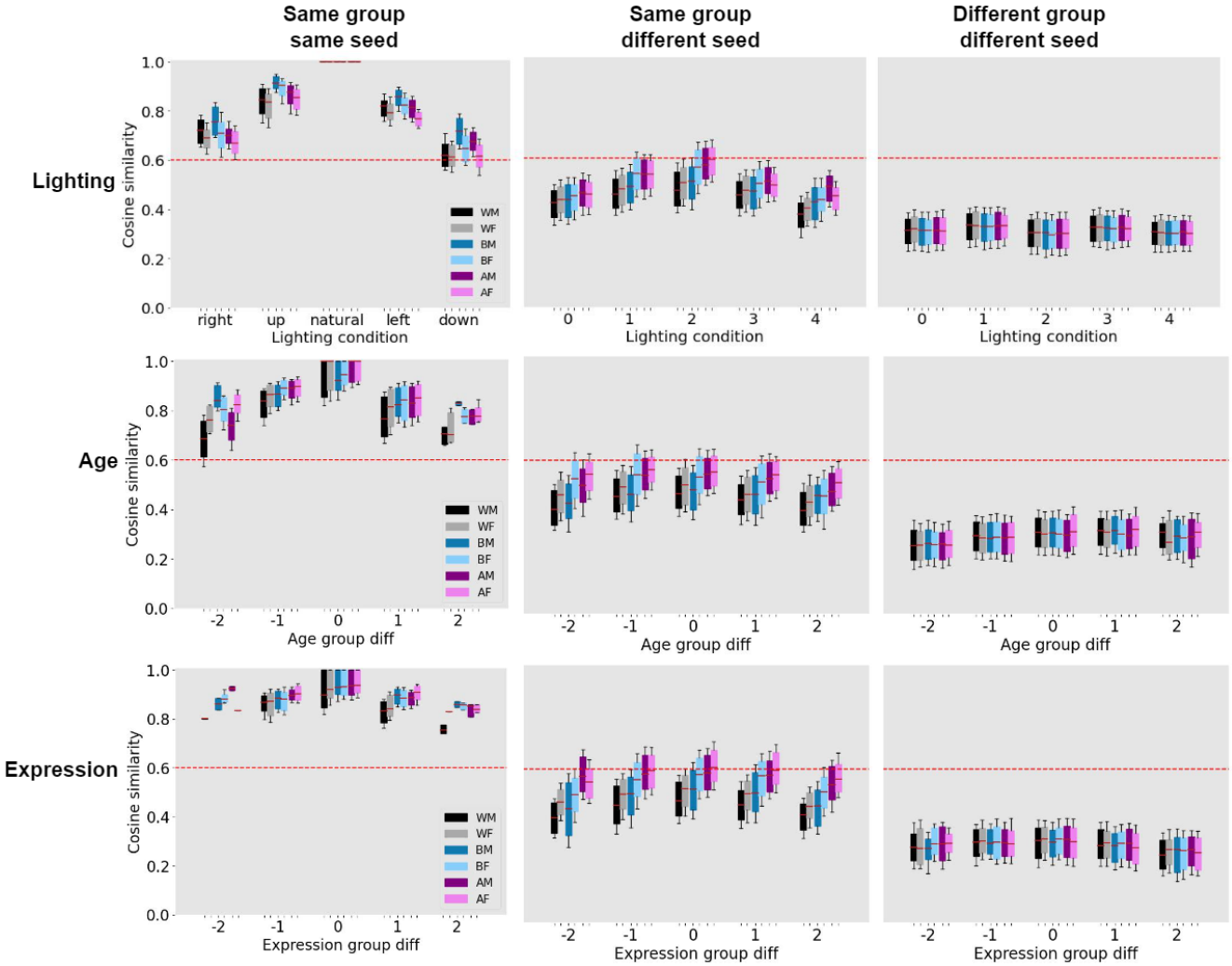

Figure 4. **Identity similarity scores reported by ResNet34 trained using ArcFace on the MS1MV3 dataset with respect to non-sensitive attributes and demographic changes**.

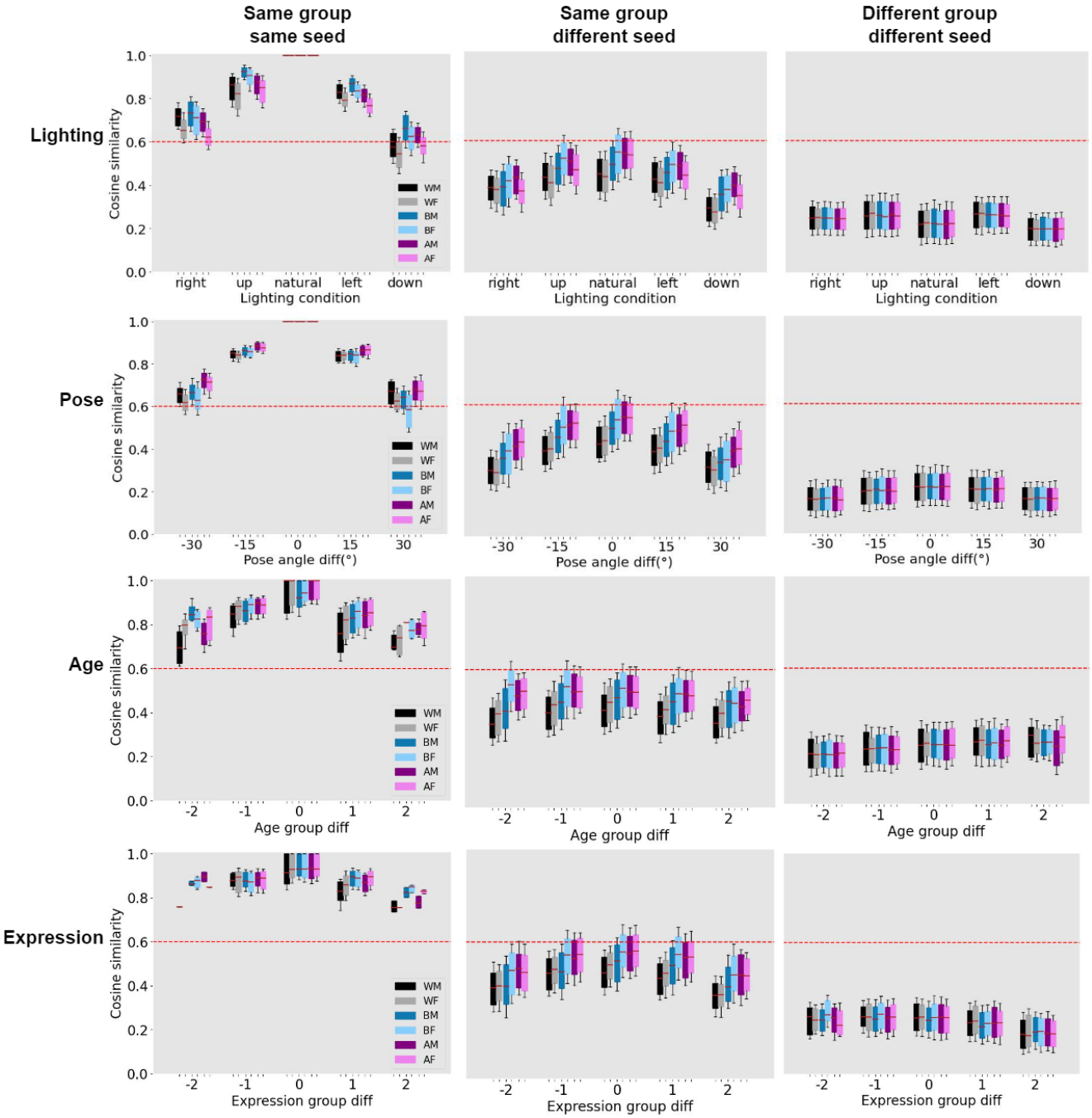

Figure 5. **Identity similarity scores reported by ResNet34 trained using ArcFace on the Glint360k dataset with respect to non-sensitive attributes and demographic changes**.

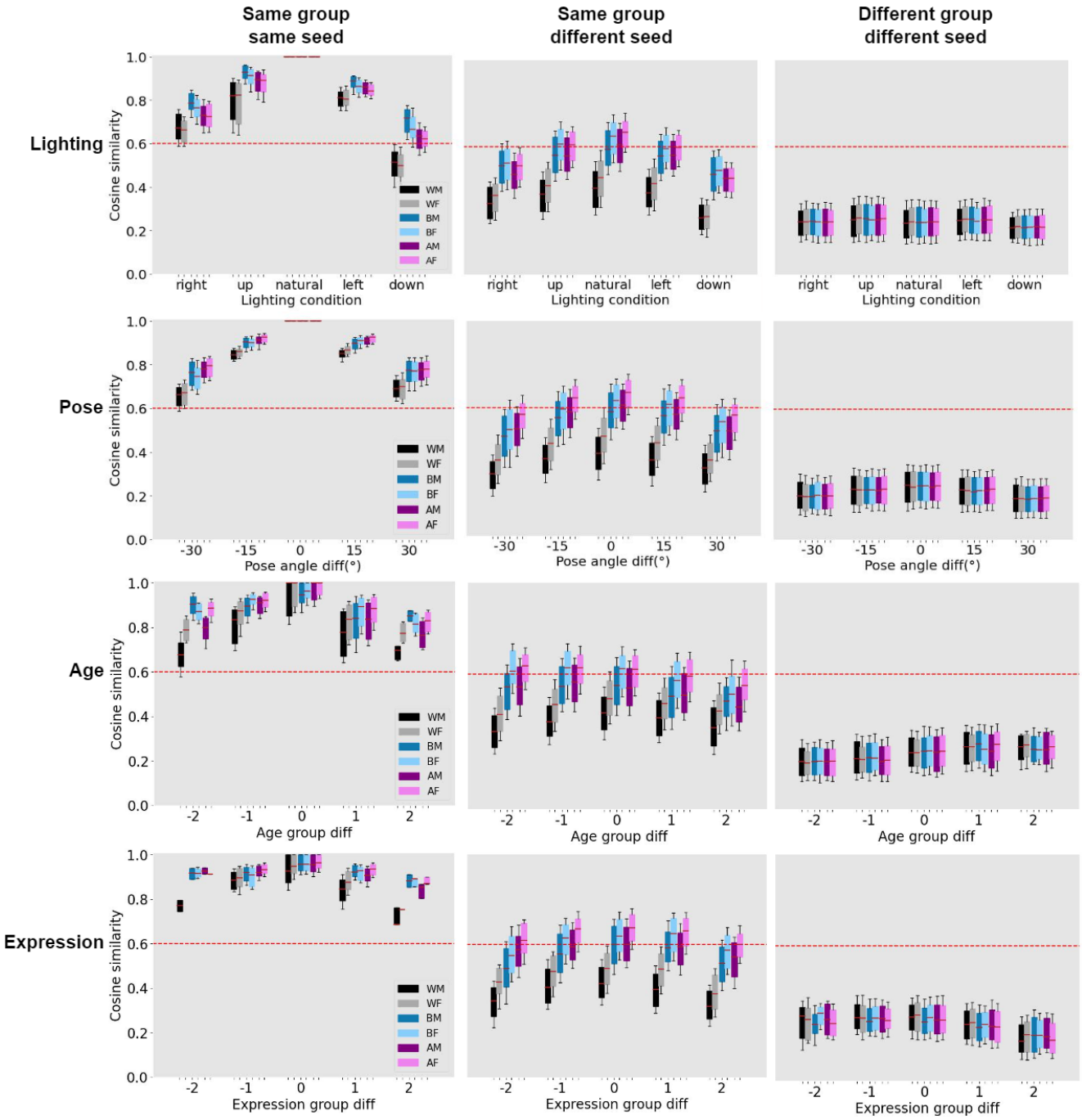

Figure 6. **Identity similarity scores reported by SFNet20 trained using SphereFace on the VGGFace2 dataset with respect to non-sensitive attributes and demographic changes**.

F. Per-Image Standard Deviations of Human Annotations

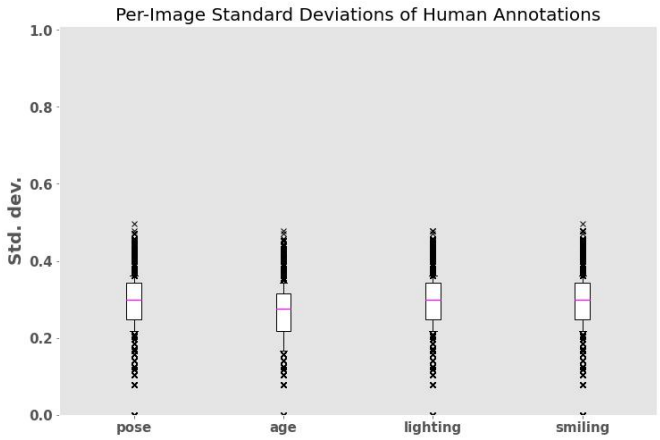

Figure 7. **Per-Image Standard Deviations of Human Annotations**. Distributions of per-image standard deviations of human annotations for each of the attributes we considered (one unit = dynamic range of the attribute). Nine annotators were asked to provide a rating for an face image pair of each image. The median standard deviations are plotted in red lines, all of the medians are around 0.3, indicating good consistency among annotators.

G. Human annotation results for different attributes

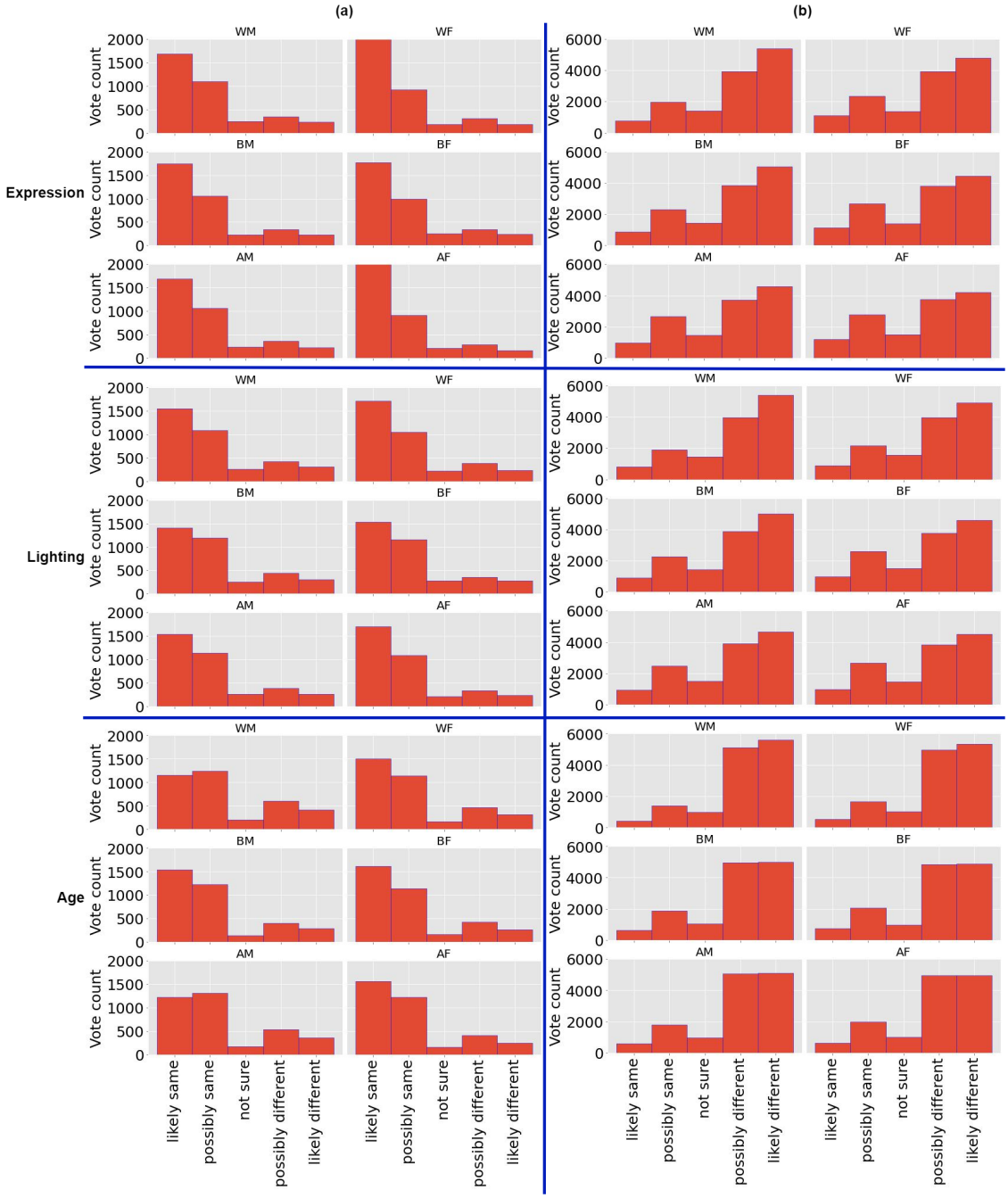

Figure 8. **Human annotation results on different attribute image pairs. (a)** Image pairs from same seed and same demographic group. **(b)** Image pairs from different seeds but same demographic group. We see a consistent trend across all attributes. Refer to Sec. 4.2 for details.

H. FNMR v.s. FMR results for different t_{hcic} values

We show results of the FNMR v.s. FMR plot with thresholds $t_{hcic} \in \{0.2, 0.4\}$ in Fig. 9, 10. The basically show the same trend as $t_{hcic} = 0.3$. Refer to Sec. 4.1.1 for details.

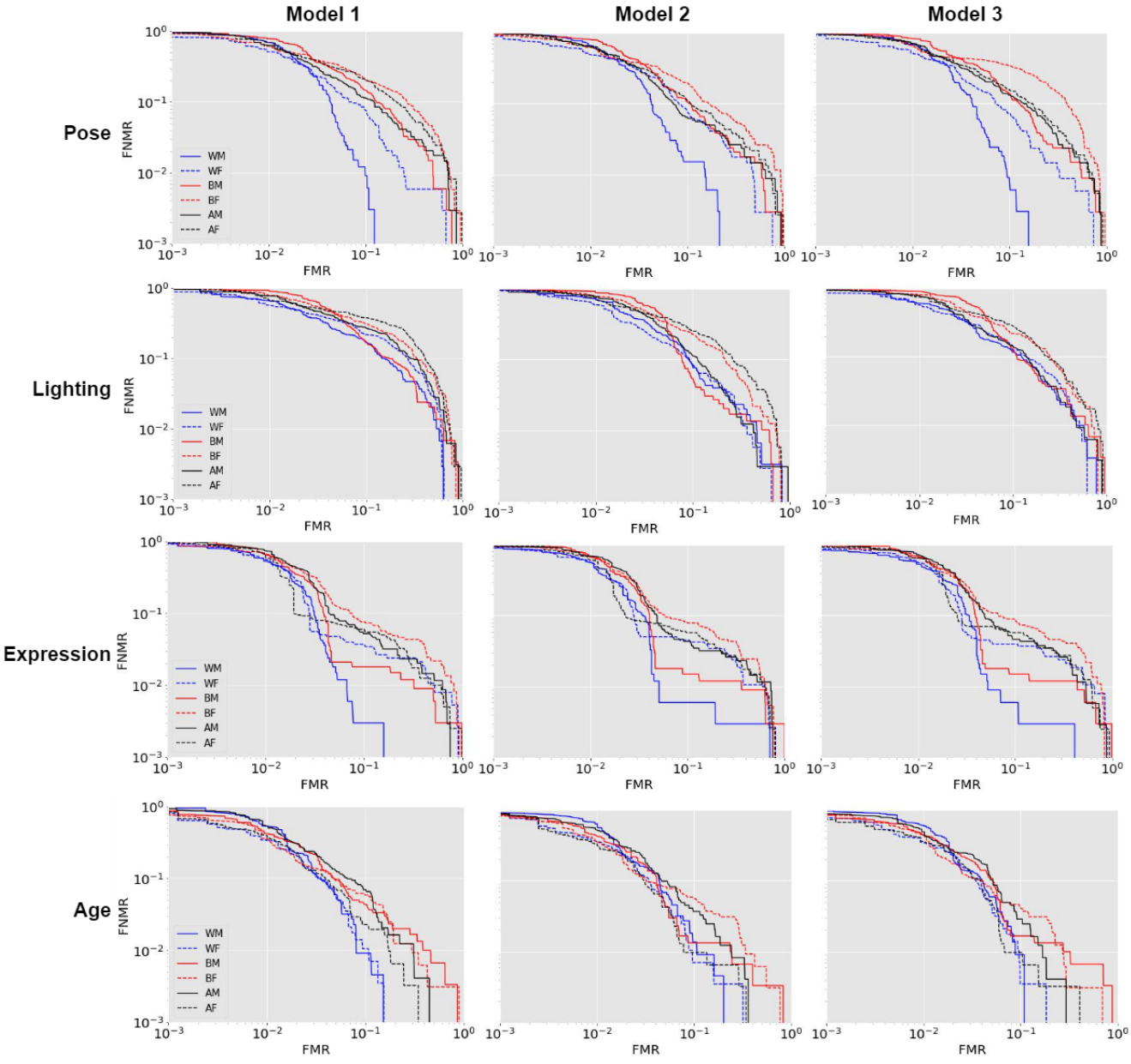

Figure 9. **FNMR vs. FMR plots** from all image pairs using HCIC with $t_{hcic} = 0.4$ as the ground truth labels.

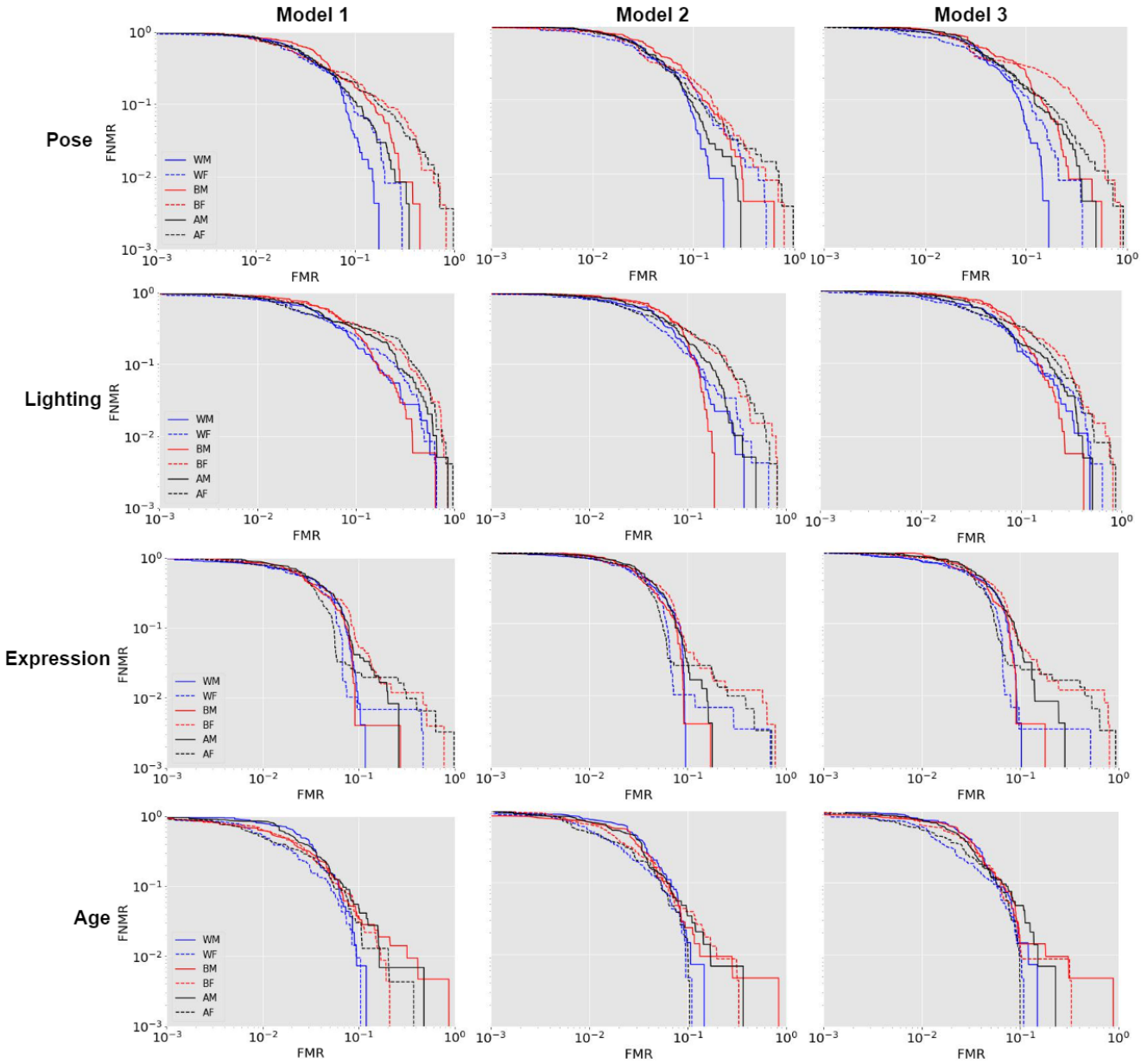

Figure 10. **FNMR vs. FMR plots** from all image pairs using HCIC with $t_{hcic} = 0.2$ as the ground truth labels.

I. More examples & failure cases

I.1. Examples of human identity evaluations for different face pairs

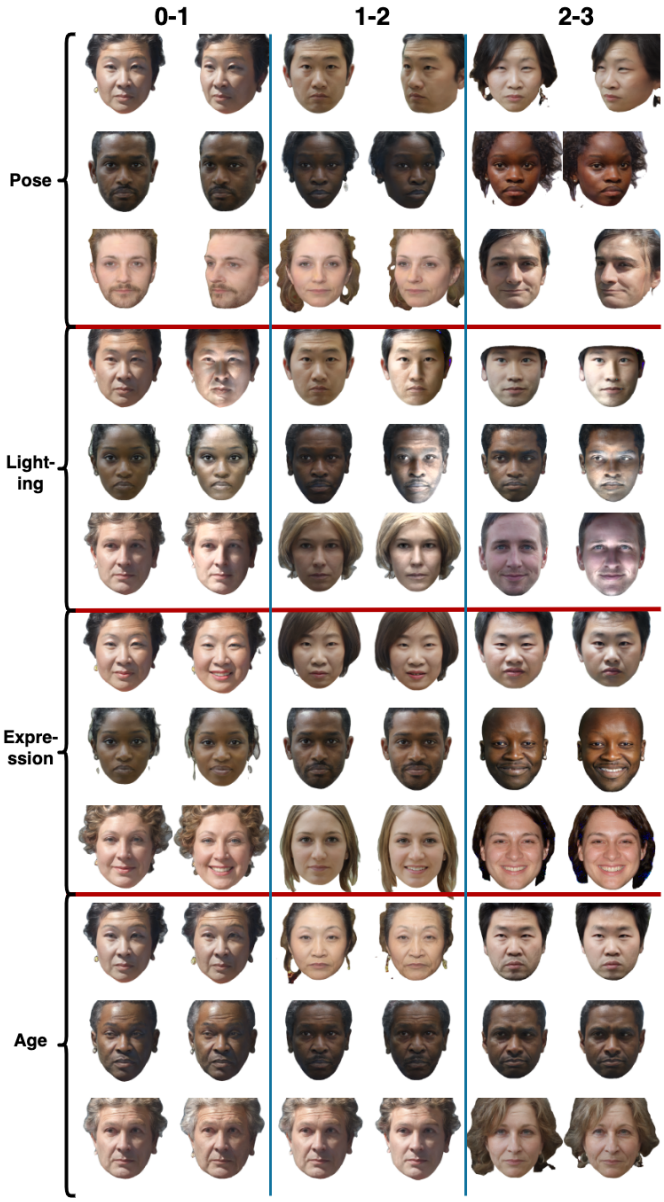

Figure 11. **Human identity annotation scores for face pairs intended to belong to the same ID**. All of the shown image pairs are from the same prototype but with non-sensitive attributes modified to a different degree, as described in Sec. 3. We show the average human annotation scores on top (high score indicates more likely to be from different IDs, raw range is used here $(0-4)$). The last column corresponds to face pairs which humans thought were from different identities, although we intended them to depict the same identity.

I.2. Failure case with large “uncanniness” score

Figure 12. **Examples of failure case**. Examples of images whose “uncanniness” score are ≥ 0.8. There are four main reasons: (a) First row, males with typically female hairstyles. (b) Second row, ”ring” artifacts. (c) Third row, bad foreground/background separation makes the hair look unrealistic. (d) Fourth row, other human subjective reasons. We remove all uncanny examples from the test database so that they will not influence experimental results.