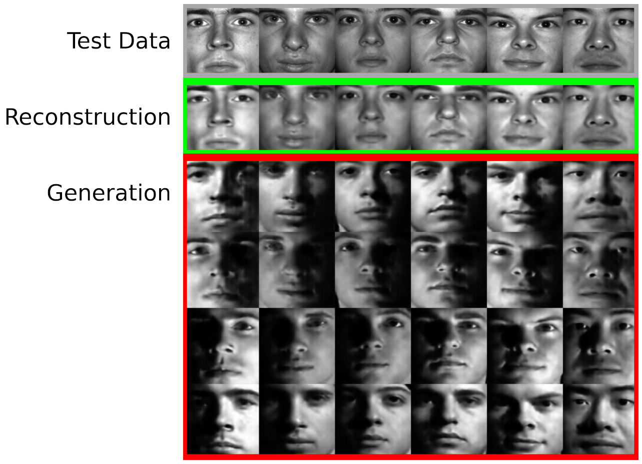
1 ADDITIONAL FIGURES

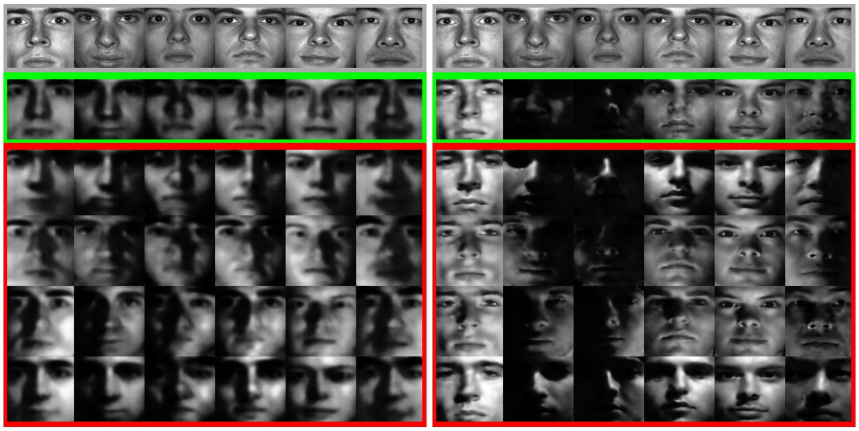
Extended Yale B Fig. 1 depicts the class-preserving samples generated from WFAE (Panel a), HCV (Panel b), and FairDisCo (Panel c) trained with the Extended Yale B dataset, which enlarges the corresponding results presented in the main figure (Figure 4 in the main script), panel A. The green box highlights generated images by encoding the test image X and nuisance data S into $Z_1 = g_2(f_1(X, *, S), f_2(X, *, S))$, and then computing $g_1(Z_1, S)$. Those in the red box were generated by using the same Z_1 but setting $S = (\pm 0.3, \pm 0.3)$. Note that WFAE has better generation quality than HCV, while keeping the identity of the input test data.

MNIST Fig. 2 depicts the class-preserving samples generated from WFAE (Panel a) and HCV (Panel b) trained with MNIST dataset, which is fuller than the corresponding results presented in the main figure, panel A. Note that all input test data were not used in training, and we did not provide information about actual digit class S of the test data on WFAE. The decoded samples are from g with estimated S and i) not using f_2 (blue box), ii) using encoded $Z = f_2(X, S)$ from the test data (green box), and iii) using Z sampled from prior P_Z (red box). For WFAE, decoded images with the same S all retained their digit information. Reconstruction without using f_2, although recognizable, produced degraded images. FairDisCo could generate images with the given S, but the generated images lacked variability compared with WFAE. HSIC constrained (conditional) variational autoencoder with the similar structure failed to generate digits of different styles. Fig. 3 presents the full result of the style transfer task shown in the main figure, panel B. We estimated Z from the source and S from the target and generated new images by $g(Z, S)$. While samples of both methods seems to successfully inherit S and Z from different target and source images, WFAE generated more diverse and clearer samples. We could infer from the resulting transfer that Z represents writing styles such as stroke thickness or degree of rotation.

VGGFace2 Fig. 4 shows the class-preserving generation results for the VGGFace2 dataset, extending the corresponding results in the main figure, panel A. Although images of persons who were *not* in the training data were used, the WFAE could successfully generate images retaining the identity while employing other identity-invariant features, e.g., camera angle, lighting condition. Fig. 5 shows the extended style transfer results corresponding to the main figure, panel B. The generated images possess the styles from the source data and tend to preserve the specified attribute of the target data. For example, the generated images tend to have open mouth if the target image has mouth wide open. Figs. 6 and 7 extend the attribute manipulation results (panel C of the main figure), including gender switch. Manipulating the gender attribute changed eyes and lips of the test data so that the images resemble stereotypical male (or female) pictures. Changing the beard attribute could draw images with or without facial hair from the source image, even for female images. Letting the sunglasses attribute positive produced decoded images having darkened eye area that resembles sunglasses; making it negative on an image with sunglasses produced one without them. Manipulating mouth open could make the manipulated images with either mouth wide open or closed. For the Fader Networks, we had to extrapolate the attribute scores to a large magnitude as far as ± 400, but it also caused excessive distortion on the original image, as shown in Fig. 7.



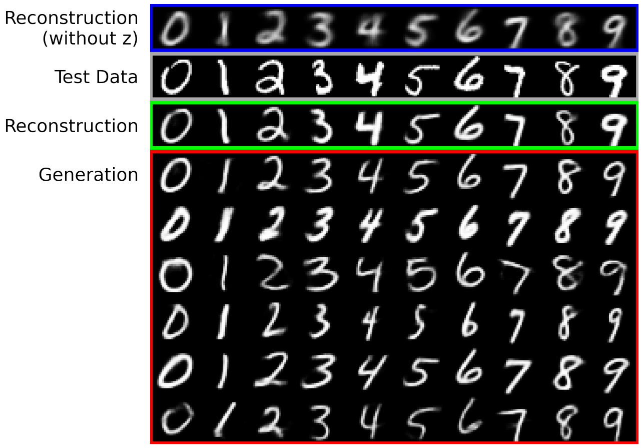
(a) WFAE



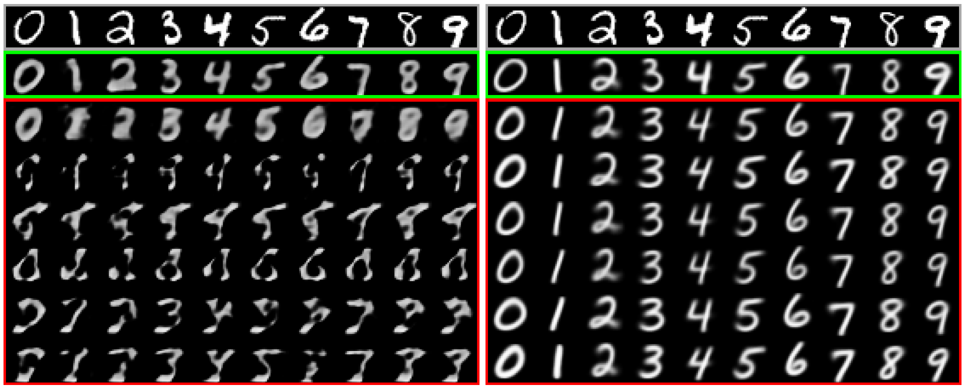
(b) HCV

(c) FairDisCo

Figure 1: Class-preserving generation from the Extended Yale B dataset.



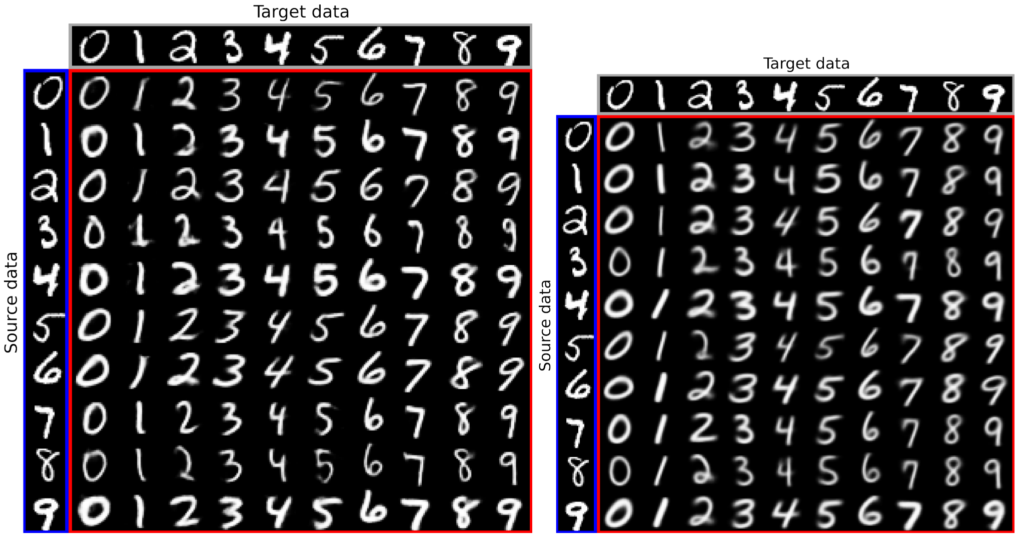
(a) WFAE



(b) HCV

(c) FairDisCo

Figure 2: Class-preserving generation from the MNIST dataset.



(a) WFAE

(b) FairDisCo

Figure 3: Style transfer in the MNIST dataset.

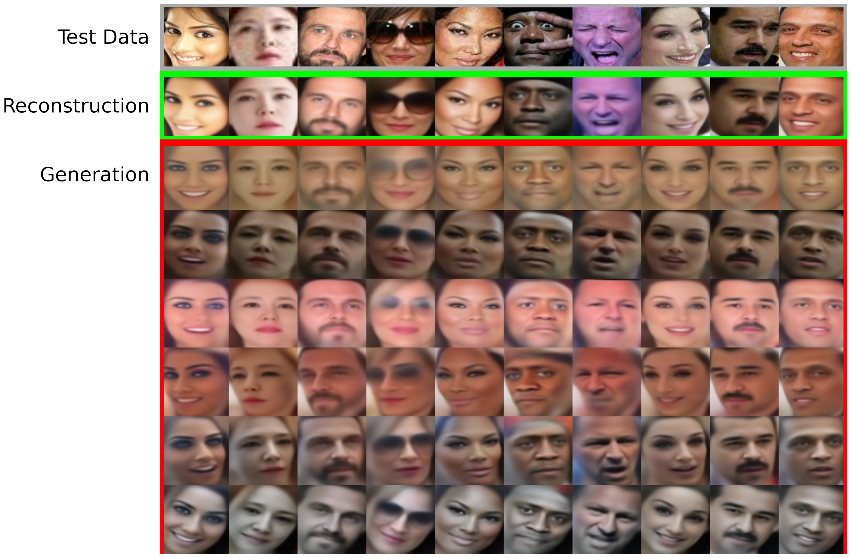


Figure 4: Class-preserving generation from the VGGFace2 dataset.

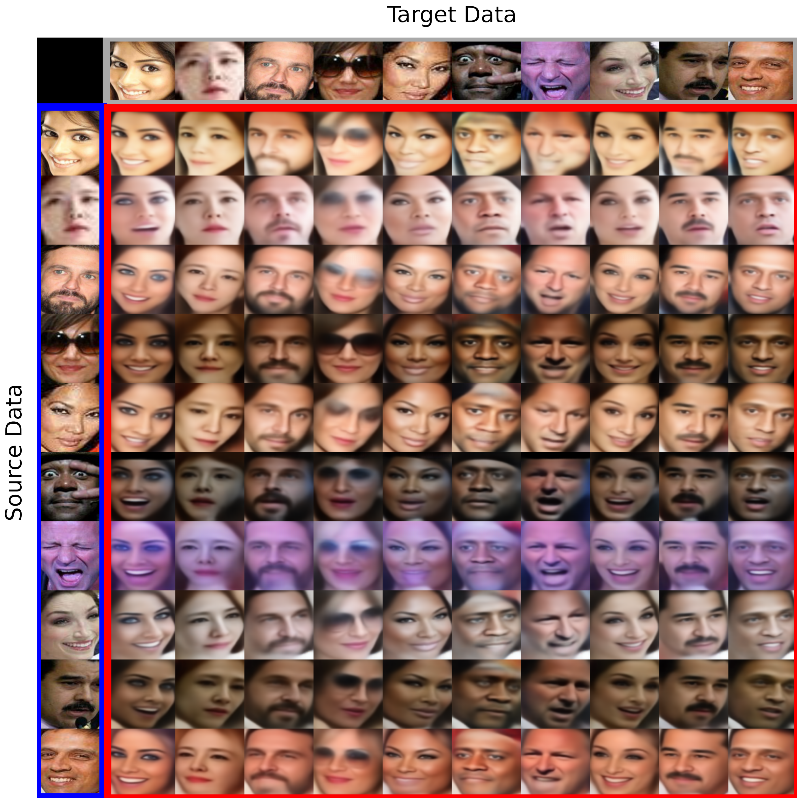
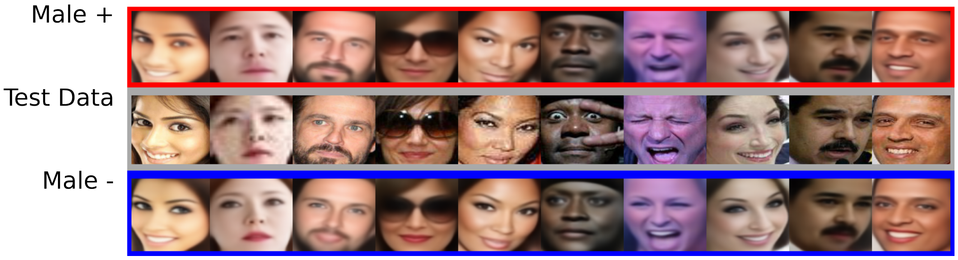
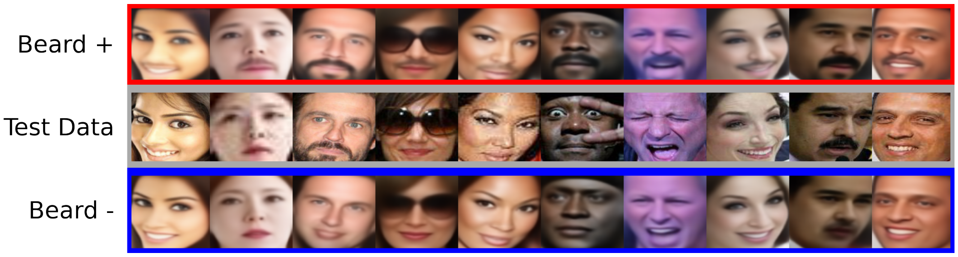


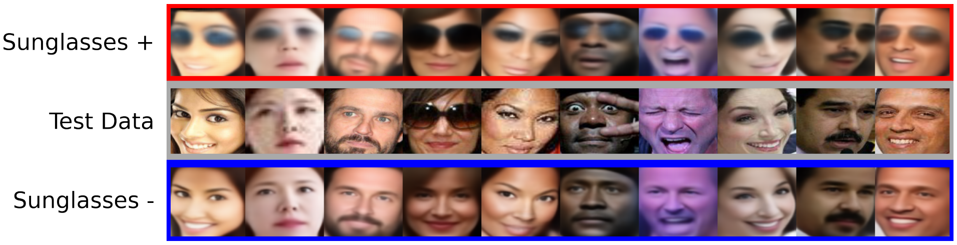
Figure 5: Style transfer in the VGGFace2 dataset.



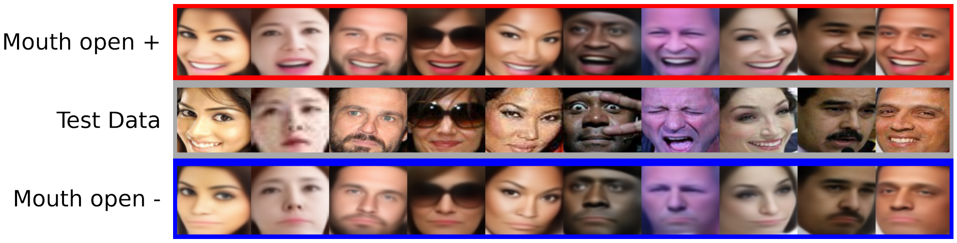
(a) Attribute Manipulation - Male



(b) Attribute Manipulation - Beard

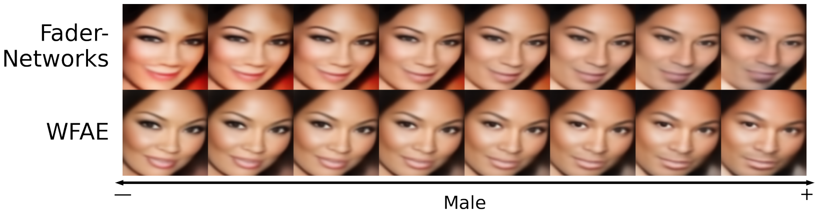


(c) Attribute Manipulation - Sunglasses

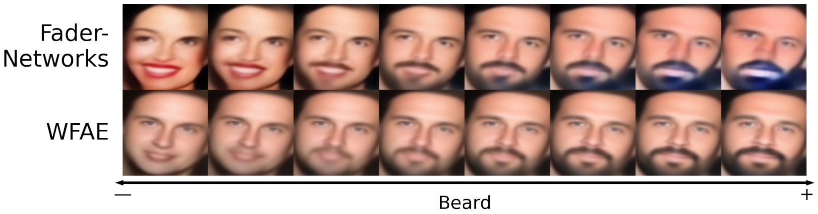


(d) Attribute Manipulation - Mouth open

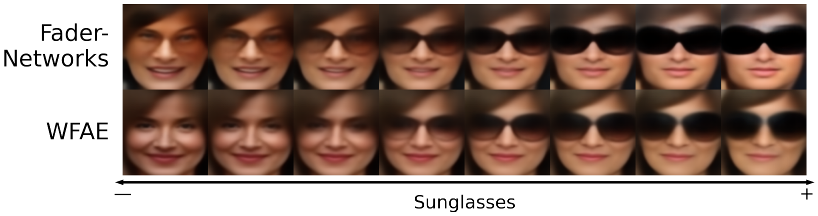
Figure 6: Decoded images with attribute score manipulated to either 4.0 (red box, first row) or -3.0 (blue box, third row).



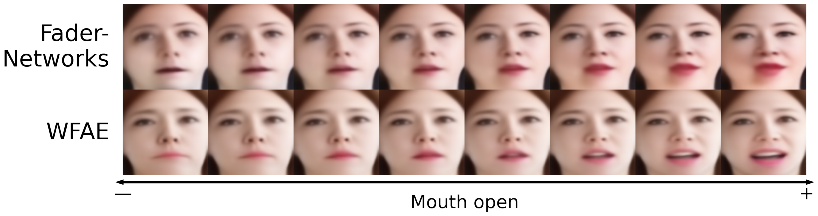
(a) Attribute Interpolation - Male



(b) Attribute Interpolation - Beard



(c) Attribute Interpolation - Sunglasses



(d) Attribute Interpolation - Mouth open

Figure 7: Decoded images with attribute score interpolated in trained Fader Networks and WFAE