Hierarchical Diffusion Autoencoders and Disentangled Image Manipulation Supplemental Material

Appendix: Overview

In Section A, we provide details on the architecture design and training setups. In Section B, we provide more experiments and visual results to demonstrate the effectiveness of Hierarchical Diffusion Autoencoders.

A. Appendix: Architectures

Our Hierarchical Diffusion Autoencoders adopts the same diffusion-based decoder architecture with Diffusion Autoencoders [7] [1]. The network architectures of Resblocks in the diffusion U-net and semantic encoder are shown in Fig. 1. Following Diffusion Autoencoders, the timestep embeddings and semantic codes are fed into the diffusion-based decoder as conditions with AdaGN layers. For the fairness of the experiments, we try to keep most of the hyperparameters consistent with Diffusion Autoencoders [7].

The network architecture, hyperparameters, and training parameters of Hierarchical Diffusion Autoencoders are shown in Tab. 1.

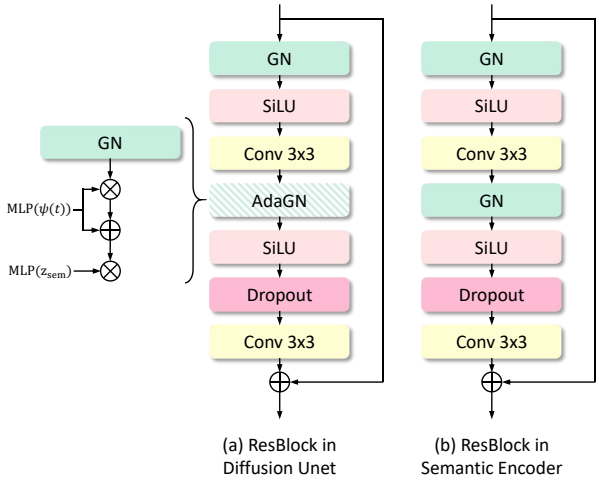


Figure 1. **Architecture overview of Hierarchical Diffusion Autoencoders.**

[1] https://github.com/phizaz/diffae

Figure 2. **The visualization of the dilated binary masks generated from the segmentation maps of the eyes.** We highlight the area to be edited, and the remaining non-highlighted areas should be consistent before and after editing in our image manipulation with high fidelity task.

B. Appendix: Experiments

B.1. Evaluating the Fidelity of Image Manipulation

An important evaluation criterion for image manipulation is the fidelity of image manipulation, *i.e.*, how well the manipulated images preserve the details of the original images.

We conduct the experiments on the face datasets CelebA-HQ [1]. Given an image I_0 and an attribute a to manipulate, we first extract a binary mask I^a_{bin} to indicate the rough regions related to the attribute, by dilating the segmentation map of the attribute. I^a_{bin} highlights the area to be manipulated ($I^a_{bin} = 1$ in this area), and $1 - I^a_{bin}$ highlights the area that should be consistent before and after editing, as shown in Fig. 5. Then we apply our Hierarchical Diffusion Autoencoders to manipulate the original image I_0 with the specific attribute a to get the manipulated image I^a_m. Finally, we can compute the LPIPS [17] and MSE metrics between $I_0 * (1 - I^a_{bin})$ and $I^a_m * (1 - I^a_{bin})$. We calculated the average value of MSE and LPIPS from $\alpha = -0.5$ to $\alpha = 0.5$ (α is the coefficient of editing direction, which controls the degree of editing) and Tab. 2 reports the results of DAE, HDAE(U) and HDAE(U) with disentangled attribute manipulation. The results demonstrate that our HDAE(U) preserves the details better than HDAE, and the disentangled manipulation with truncated features achieves the best performance in terms of

Hyperparameter	DAE (FFHQ 128)	DAE(2560) (FFHQ 128)	DAE(U) (FFHQ 128)	HDAE(E) (FFHQ 128)	HDAE(U) (FFHQ 128)
Batch size	128	128	128	128	128
Base channels	128	128	128	128	128
Channel multipliers	[1,1,2,3,4]	[1,1,2,3,4]	[1,1,2,3,4]	[1,1,2,3,4]	[1,1,2,3,4]
Images trained	130M	130M	130M	130M	130M
Encoder base ch	128	128	128	128	128
Encoder ch. mult.	[1,1,2,3,4,4]	[1,1,2,3,4,4]	[1,1,2,3,4,4]	[1,1,2,3,4,4]	[1,1,2,3,4,4]
Decoder base ch	-	-	128	-	128
Decoder ch. mult.	-	-	[1,1,2,3,4,4]	-	[1,1,2,3,4,4]
Attention resolution	[16]	[16]	[16]	[16]	[16]
Latent code dim	512	2560	512	2560	2560
β scheduler	Linear	Linear	Linear	Linear	Linear
Learning rate			1e-4		
Optimizer			Adam (no weight decay)		
Training T			1000		
Diffusion loss			MSE with noise prediction $\vec{\epsilon}$		
Diffusion var.			Not important for DDIM		
Parameters	122.59M	190.6M	160.94M	154.62M	189.15M

Table 1. **Network architecture and training hyperparameters of hierarchical diffusion autoencoder.**

Metric	LPIPS(↓)							
Attribute	**eyeglasses**	**mouth_slightly_open**	**big_lips**	**big_nose**	**mustache**	**bags_under_eyes**	**arched_eyebrows**	**gray_hair**
DAE [7]	0.23267	0.18339	0.10974	0.15190	0.18660	0.20546	0.18365	0.31655
HDAE(U)	0.22098	0.16874	0.10437	0.13695	0.18223	0.19380	0.17504	0.28918
HDAE(U) Disentangled	**0.18237**	**0.13382**	**0.07329**	**0.10872**	**0.14892**	**0.12983**	**0.13568**	**0.22260**
Metric	**MSE(↓)**							
Attribute	**eyeglasses**	**mouth_slightly_open**	**big_lips**	**big_nose**	**mustache**	**bags_under_eyes**	**arched_eyebrows**	**gray_hair**
DAE [7]	0.01105	0.00906	0.00341	0.00388	0.01060	0.00952	0.00588	0.08564
HDAE(U)	0.01011	0.00804	0.00262	0.00304	0.00823	0.00843	0.00548	0.07369
HDAE(U) Disentangled	**0.00771**	**0.00593**	**0.00126**	**0.00165**	**0.00517**	**0.00595**	**0.00389**	**0.04739**

Table 2. **Evaluation of image manipulation fidelity.** "HDAE(U) Disentangled" means disentangled image manipulation with truncated feature using HDAE(U).

fidelity.

B.2. Image Reconstruction

We show the images reconstructed from DAE [7] and HDAE(U) with their corresponding x_T as well as random x_T for comparison. We also show the results from some state-of-the-art GAN-based methods, such as HFGI [14] and PSP [8]. As shown in Fig. 6, our HDAE can achieve better reconstruction results than DAE with random x_T, indicating that our HDAE encodes richer and more comprehensive representations in the hierarchical semantic latent code.

Ablation study on number of blocks in UNet. We change the number of blocks to obtain a larger HDAE-L model (304M params) and a smaller HDAE-S model (54M params). We refer to the original HDAE in main paper as HDAE-B (190M params). We compare those models trained for 340M steps on the test set of FFHQ. The MSE of HDAE-S, HDAE-B and HDAE-L are 0.007714, 0.004847 and 0.004262, respectively. The LPIPS of HDAE-S, HDAE-B and HDAE-L are 0.09718, 0.06736 and 0.06235, respectively. Results indicate the potential to scale up our model.

B.3. Image Manipulation

Detail-preserving image manipulation. Fig. 7 shows the visual results of image manipulation on real images with GAN inversion approach HFGI [14], E4E [12], our baseline DAE [7], and our proposed model HDAE(U). Fig. 8 further compares the visual results of image manipulation on real images between DAE and HDAE(U). The qualitative results demonstrate that our HDAE(U) is extremely good at preserving details for image editing, compared with previous approaches.

Disentangled image manipulation with truncated features. We show the qualitative results of image manipulation with different α (the weight of classifier direction) and k (k channels are preserved after truncation) in Fig. 11. It

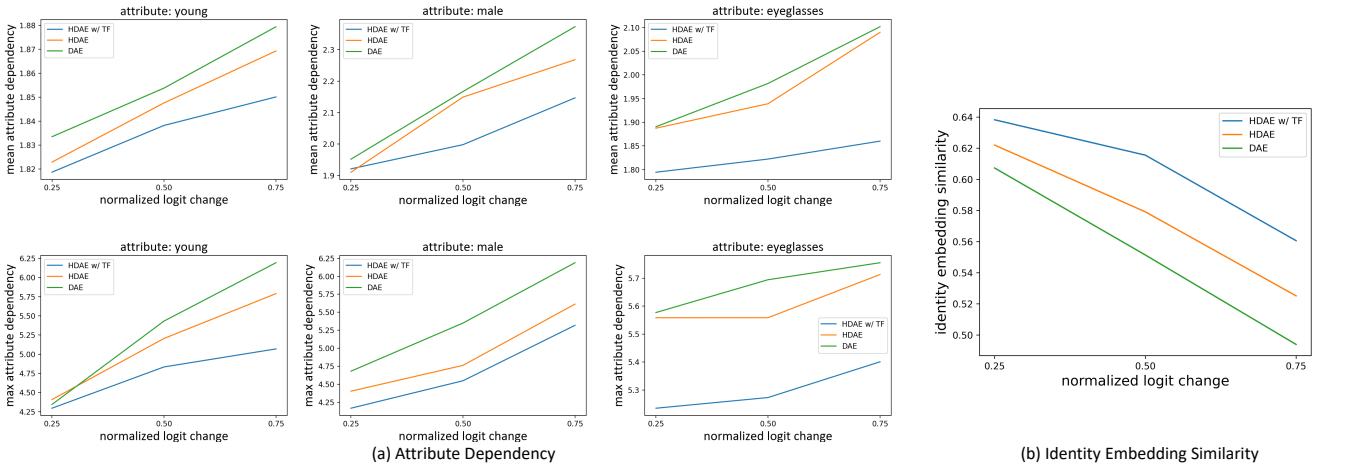


Figure 3. (a) Attribute dependency (AD) vs. the degree of target attribute manipulation. **Lower AD indicates better disentanglement**. (b) Identity embedding similarity between edited and original faces vs. the degree of target attribute manipulation. **Higher similarity indicates better identity preservation**.

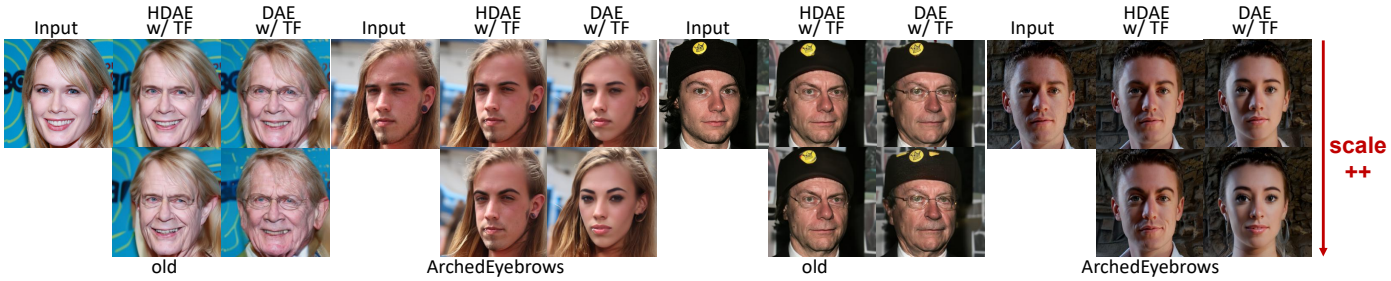


Figure 4. **Disentangled attribute manipulation results with increasing degrees of attribute manipulation.** The images in the same row are generated with the same degree of manipulation. The degree of manipulation in the second row is higher than that in the first row.

	HDAE(U)	SPADE	CLADE	GroupDNet	Pix2PixHD
FID	**23.37**	29.2	30.6	25.9	38.5

Table 3. **Image translation results of HDAE(U), SPADE [5], CLADE [11], GroupDNet [18], and Pix2PixHD [13] on the CelebAMask-HQ [3] test set.**

is shown that α controls the strength of editing and k controls the degree of disentanglement. Higher α leads to more intense editing, and lower k leads to more disentangled manipulation. We provide more qualitative results of image manipulation with truncated features in Fig. 9, demonstrating the effectiveness of our approach for disentangled image manipulation. As shown in Fig. 4 and Fig. 11, we also control the degree of attribute manipulation to compare DAE w/ TF and HDAE w/ TF.

Identity preservation for face editing. We use ArcFace to calculate the cosine similarity between the feature embeddings of 5000 pairs of original images and manipulated images. With the same degree of target attribute manipulation (normalized logit change used in [15]), HDAE w/ TF best preserves the identity, followed by HDAE, while DAE preserves the worst, as shown in Fig. 3 (b).

Disentanglement Metric: Attribute Dependency. Attribute dependency is a disentanglement metric proposed by [15] which measures the degree to which manipulating an attribute introduces changes in other attributes. Following [15], we measure the mean-AD and max-AD, as shown in Fig. 3 (a). HDAE w/ TF gets the lowest AD, followed by HDAE, while DAE gets the highest.

B.4. Interpreting the Hierarchical Latent Space

Style mixing. We show more qualitative results of style mixing in Fig. 10.

Image interpolation with different latent codes. We show the qualitative results of image interpolation with different latent codes in Fig. 13. It shows a clear hierarchy of the latent space. The low-level features control the spatial details such as background, color, and lighting, and high-level features control the global and abstract semantic attributes related to image structure such as pose, gender, face shape, and

(a) Face

(b) Horse

Figure 5. **Unconditional samples (uncurated) from our HDAE-M and latent DDIM trained on FFHQ-128 and LSUN horse-128.**

eyeglasses.

Visualization of empirical cumulative distribution function and values of the 5 x 512-dimensional normalized classifier weights. We show more examples of empirical cumulative distribution function and values of the 5 x 512-dimensional normalized classifier weights in Fig. 14.

Connections between the hierarchical latent space and truncation-based approach. The hierarchical latent space and truncation-based approach are orthogonal approaches to improve image manipulation from different perspectives. The feature hierarchy improves image manipulation by providing a comprehensive and semantically meaningful latent space. The richness of the latent space ensures detail preservation for image manipulation. On the other hand, the truncation-based approach empowers disentangled image manipulation. Therefore, HDAE(U) with truncated features shows the best detail-preserving and disentangled image manipulation results.

Dataset	Model	FID(↓) T=10	T=20	T=50	T=100
CelebA 64	DAE*	**12.92**	10.18	7.05	5.30
	HDAE-M(E)	13.19	**9.86**	**6.63**	**5.13**
FFHQ 128	DAE*	21.24	**17.15**	**13.08**	**10.93**
	HDAE-M(E)	**20.73**	17.36	14.24	12.66
Horse 128	DAE*	12.60	10.23	8.57	**8.02**
	HDAE-M(U)	**11.29**	**9.79**	**8.39**	8.22

Table 4. **Unconditional image generation results of DAE [7] and HDAE-M(E) on the CelebA, FFHQ and LSUN Horse.** $*$ denotes results produced by our re-implementation.

B.5. Details of human perceptual experiments.

We collect votes from 15 participants for our human perceptual experiments. Each participant answers 35 three-choice questions for image manipulation experiment, 35 four-choice questions for disentangled manipulation experiment, and 15 two-choice questions for image reconstruction experiment. We show some examples of our human perceptual experiments in Fig. 15.

B.6. Semantic Image Synthesis

We use HDAE(U) to transform semantic layouts into realistic images. To fully leverage the semantic information, the semantic label map is injected into the semantic encoder pretrained for semantic layouts to obtain semantic vectors z_s. The stochastic code x_T is a randomly sampled Gaussian noise map. Our HDAE(U) is trained on the CelebAMask-HQ [3] dataset with image sizes of 256×256. Tab. 3 reports the results of HDAE(U), SPADE [5], CLADE [11], GroupDNet [18], and Pix2PixHD [13]. As shown in Fig. 12, HDAE(U) can produce a superior performance on fidelity and learned correspondence without any special design for this task. By sampling different Gaussian noise maps x_T, the model can produce diverse high-quality images with the same layout.

B.7. Unconditional Image Generation

Methods. We conduct the unconditional image generation experiments on CelebA [4], FFHQ [2], and LSUN Horse [16]. Since the dimension of our hierarchical semantic vectors is higher than the dimension of the semantic vector of DAE, it is more difficult for HDAE than DAE to predict the latent semantic vectors with a latent DDIM. So we use a linear layer to map our hierarchical semantic vectors from HDAE encoder into a 512-dimensional vector. And we use this 512-dimensional vector as the condition of the diffusion U-Net network. We denote the model as HDAE-M. After

tuning (or training from scratch) our HDAE-M few epochs, we train a latent DDIM to generate the semantic latent code from random noise.

Experiments. We compute FID scores between $50,000$ randomly sampled real images from the dataset and our $50,000$ generated images. Tab. 4 reports our experiments on CelebA with image size 64×64, FFHQ of image size 128×128, and LSUN Horse of image size 128×128. Our approach performs better than DAE on two out of three datasets. We show more qualitative results of unconditional image generation in Fig. 4.

B.8. Limitations

HDAE models are trained on same-category images such as face images, while future work can explore complex scenes. Moreover, it is worth exploring whether HDAE can be generalized to pretrained text-to-image diffusion models such as Stable Diffusion [10].

References

[1] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *NeurIPS*, 2018. 1

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 4

[3] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 3, 4

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 4

[5] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 3, 4

[6] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 8

[7] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 1, 2, 4, 7, 8, 9

[8] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *CVPR*, 2021. 2, 7, 12

[9] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *TOG*, 2021. 7, 8

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 5

[11] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Efficient semantic image synthesis via class-adaptive normalization. *TPAMI*, 2022. 3, 4

[12] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *TOG*, 2021. 2, 7, 8

[13] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018. 3, 4

[14] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity GAN inversion for image attribute editing. In *CVPR*, 2022. 2, 7, 8

[15] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *CVPR*, 2021. 3

[16] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *arxiv:1506.03365*, 2015. 4

[17] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1

[18] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *CVPR*, 2020. 3, 4

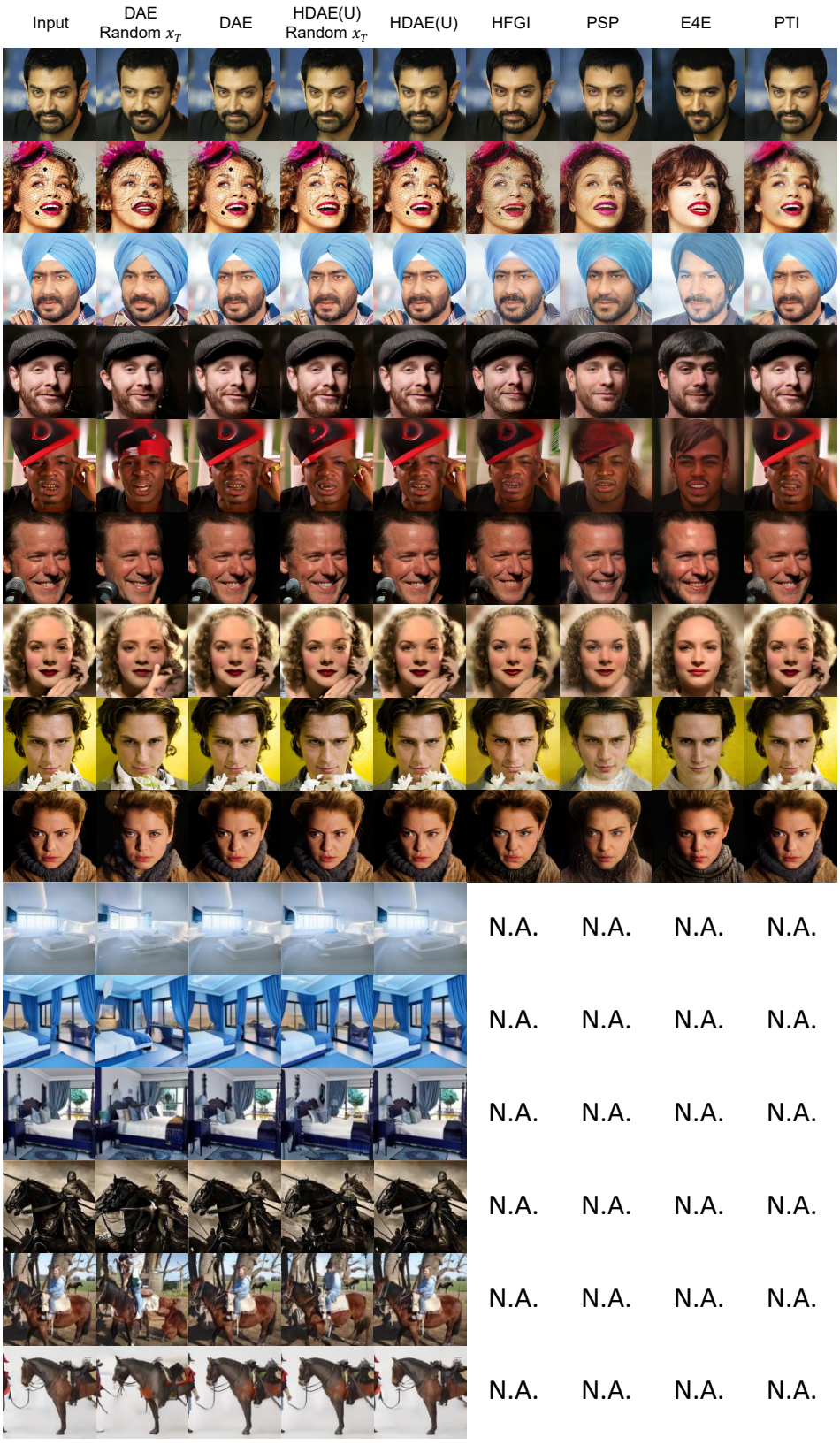


Figure 6. **Quantitative results of face and cat image reconstruction between HFGI [14], PSP [8], E4E [12], PTI [9], DAE [7] and HDAE(U).**

Figure 7. **Comparisons on real image manipulation between HFGI [14], E4E [12], PTI [9], StyleCLIP [6], DAE [7] and HDAE(U).**

Figure 8. **Comparisons on real image manipulation between DAE [7] and HDAE(U).**

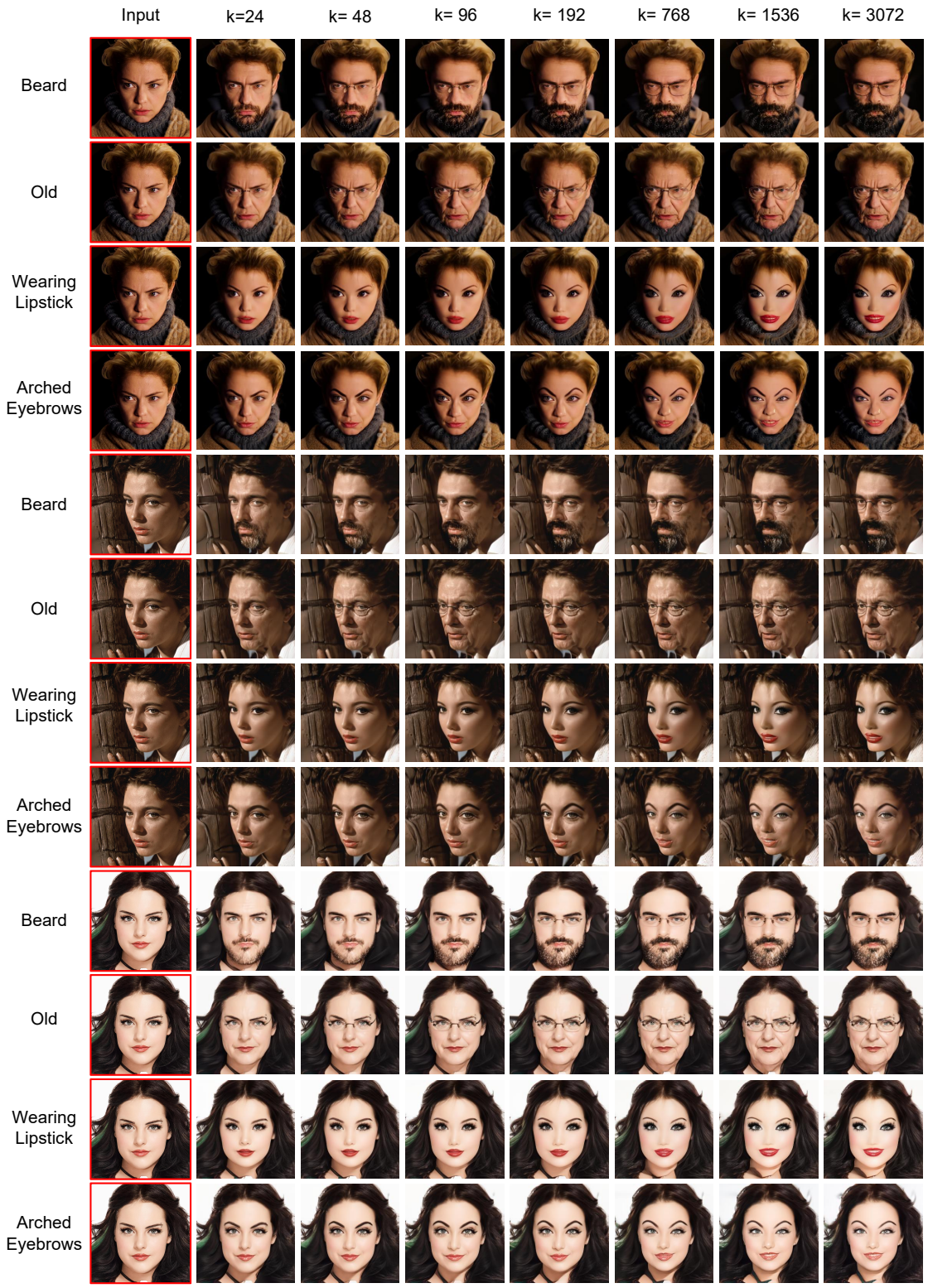


Figure 9. **Disentangled attribute manipulation results.**

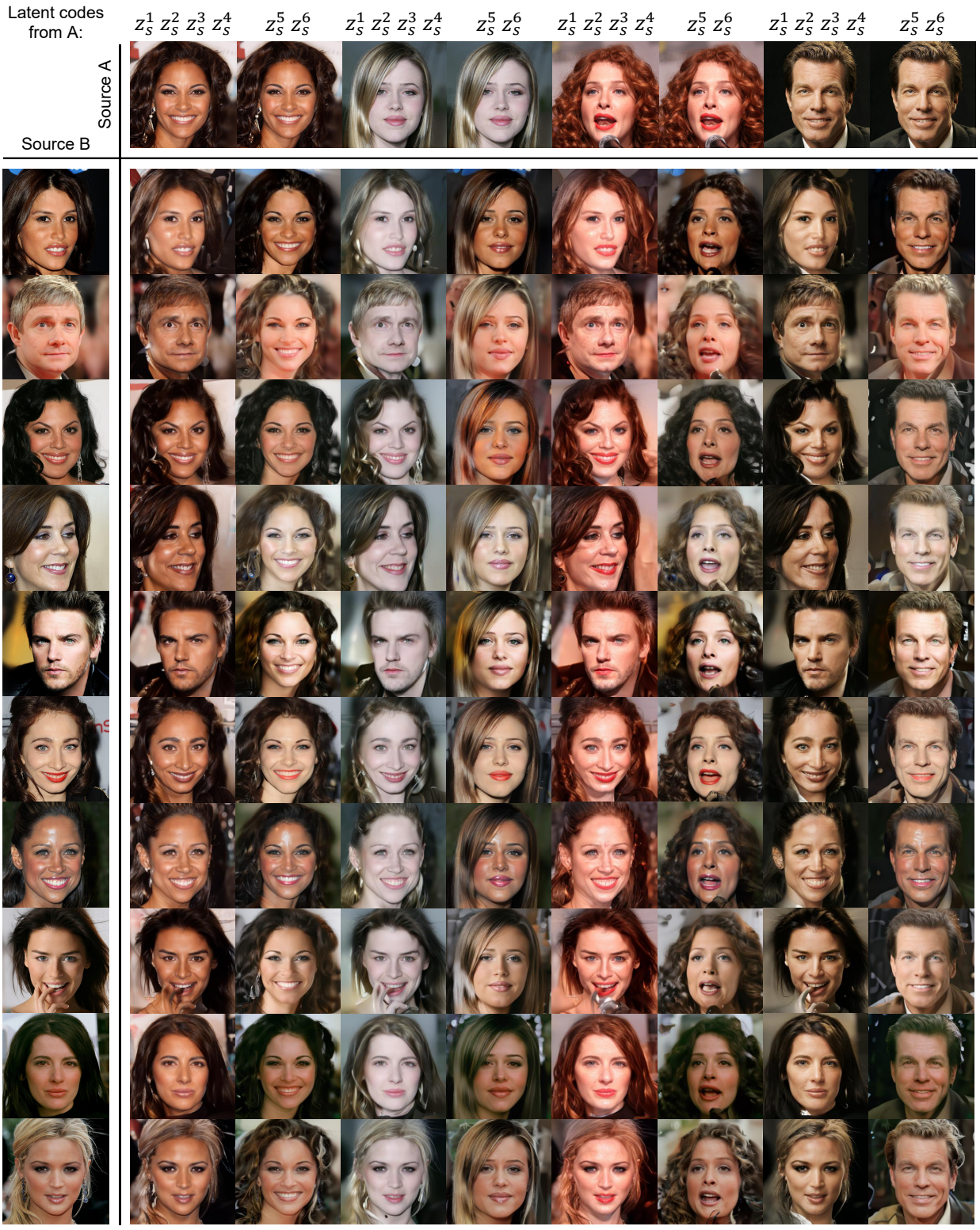


Figure 10. **Style mixing results with hierarchical latent space.**

Figure 11. **Disentangled attribute manipulation results with different α and k.** As shown in the Figure, we manipulate the attribute of beard. And we can see that glasses will appear, as k increases,

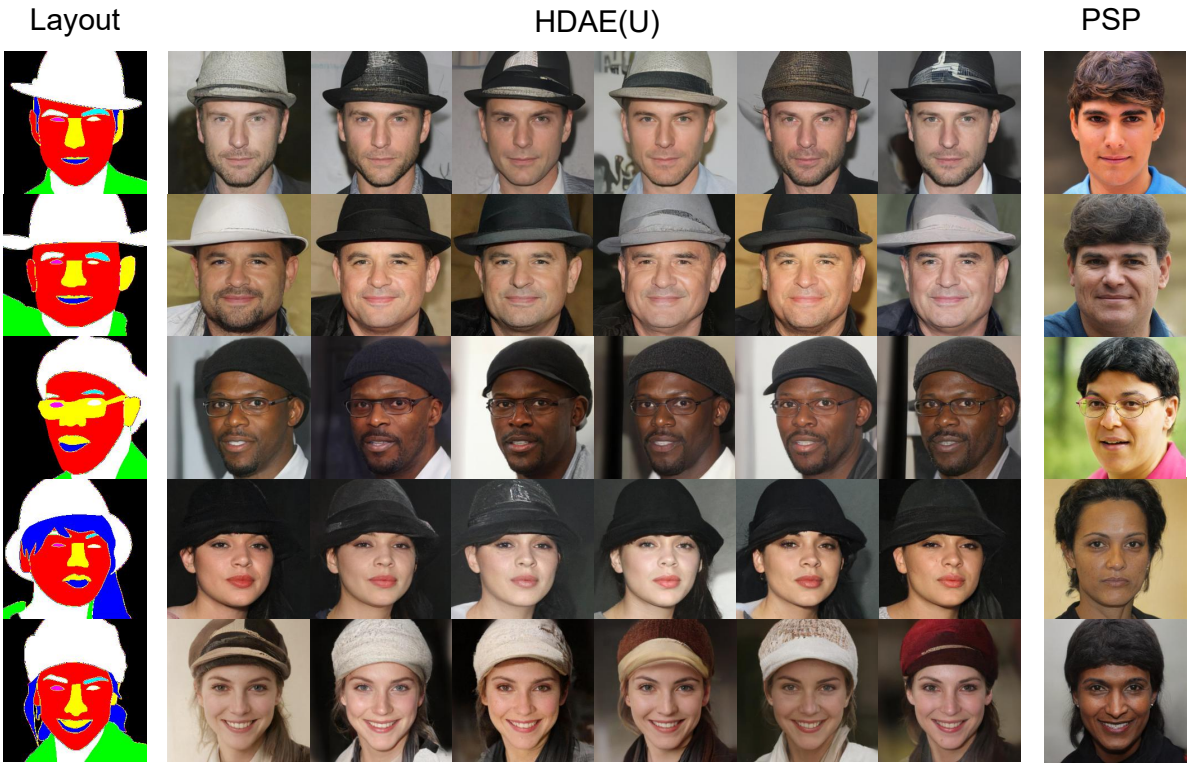


Figure 12. **Semantic image synthesis results between HDAE(U) and PSP [8]**

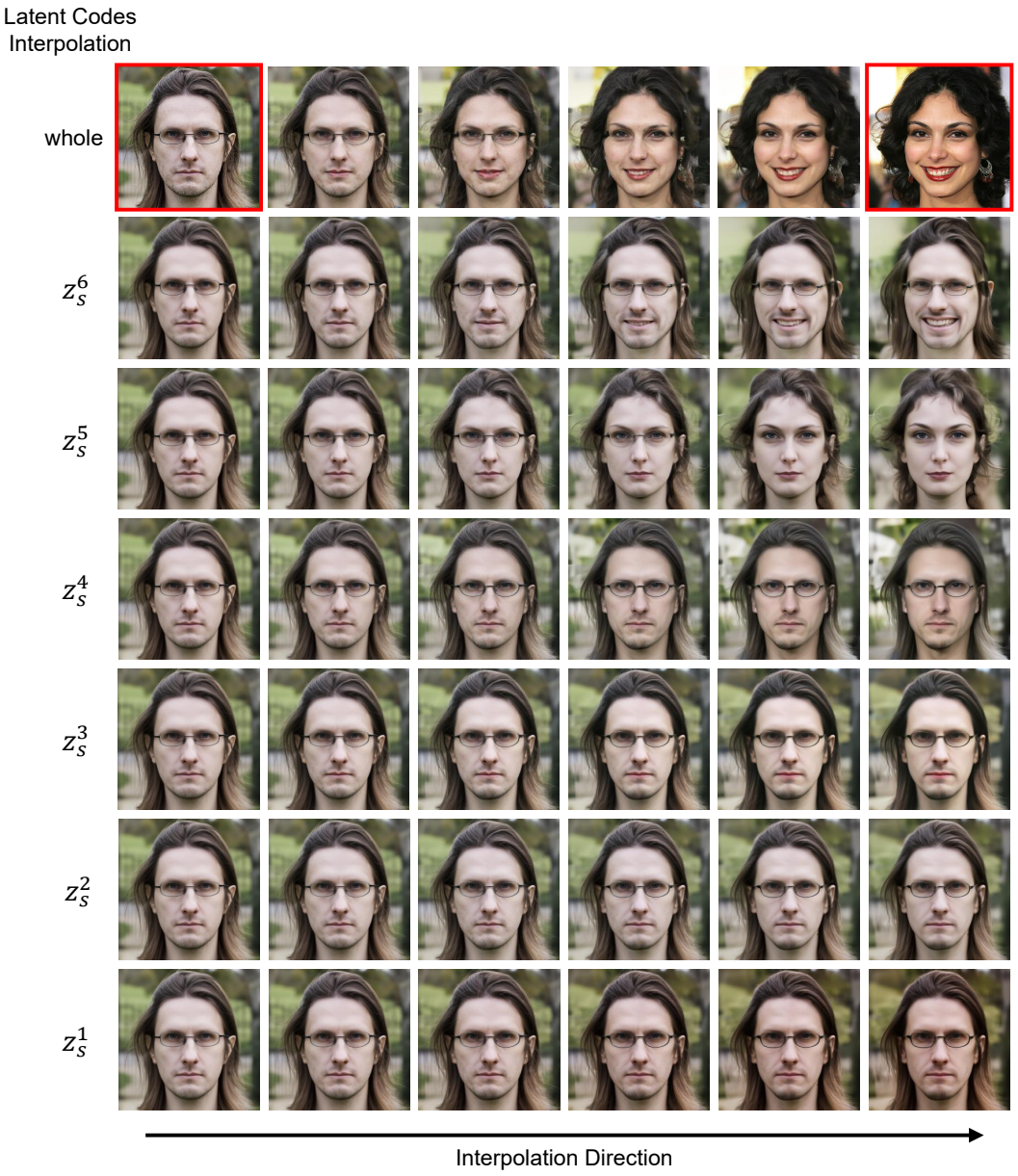
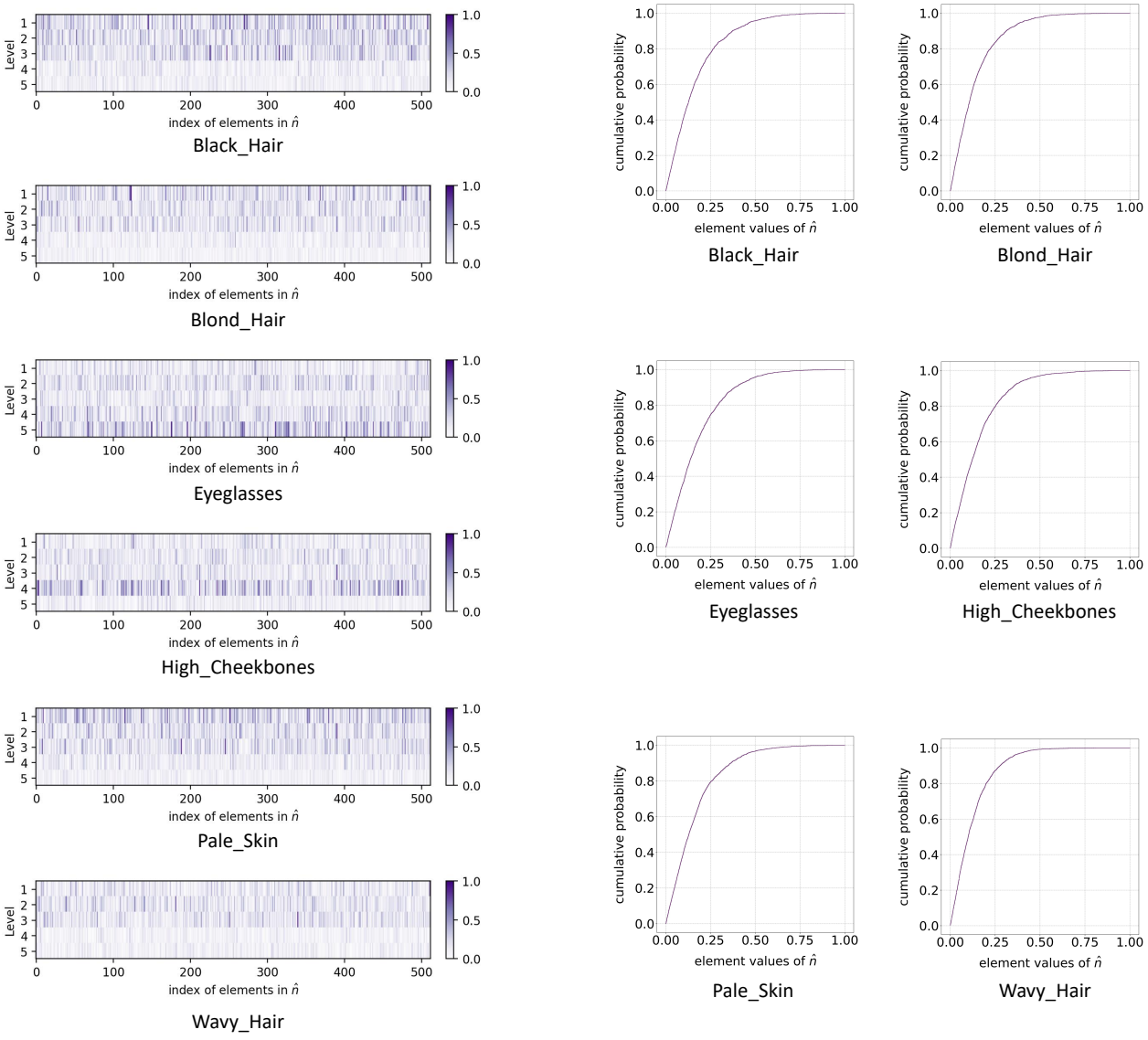


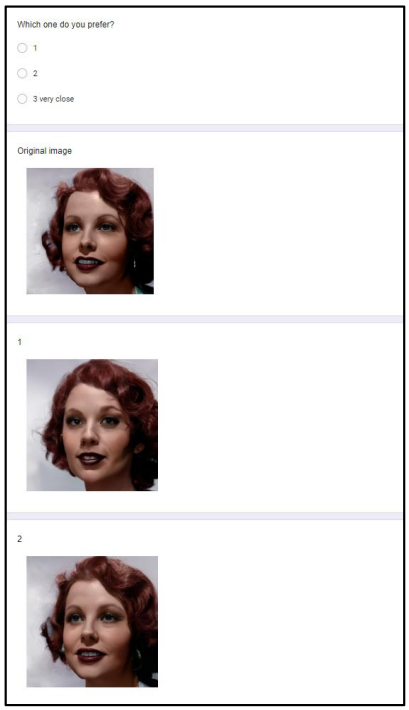
Figure 13. **Image interpolation results with different latent codes.**



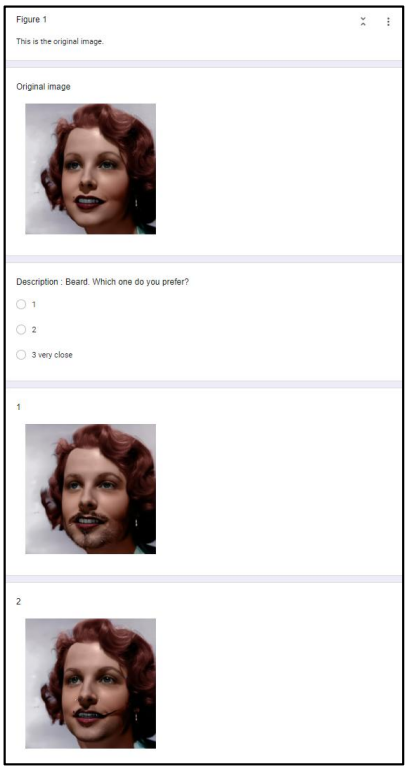
(a) values of the 5 x 512-dimensional normalized classifier weights

(b) empirical cumulative distribution function

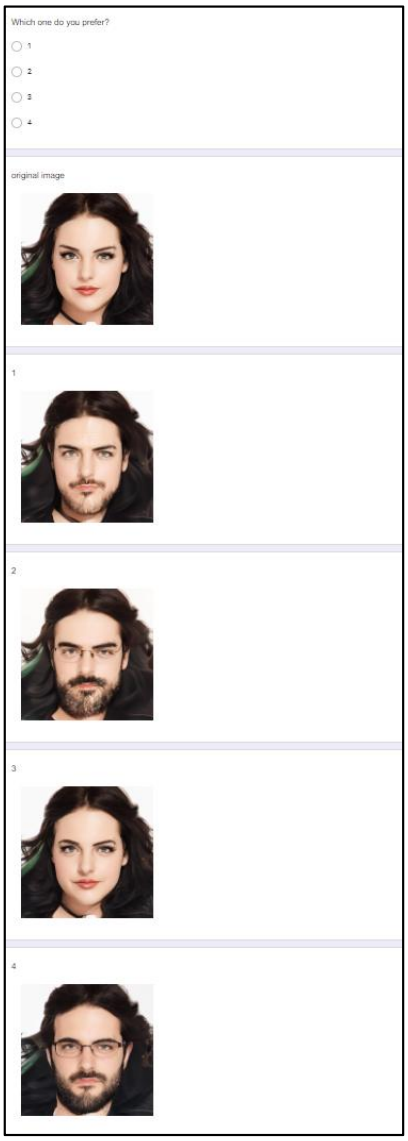
Figure 14. **More examples of the values of the 5×512-dimensional $\hat{n}$, visualized by levels and the empirical cumulative distribution function of the element values in the normalized classifier weights $\hat{n}$.**



(a) Image Reconstruction



(b) Image Manipulation



(c) Disentangled Image Manipulation

Figure 15. **User study examples of image reconstruction, image manipulation and disentangled image manipulation.**