StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On

Supplementary Material

A. Implementation details

Architecture Details. We adopt the autoencoder and the denoising U-Net of the Stable Diffusion v1.4 [4]. We initialize our denoising U-Net with the weights of the U-Net from the Paint-by-Example [5]. The U-Net's encoder and decoder both consist of 12 blocks, involving three downsampling and upsampling steps each. As a result, when StableVITON receives the input image of 64 × 48 resolution (after going through the autoencoder), three intermediate feature maps are generated for each of the following resolutions: 8×6, 16×12, 32×24, 64×48. We use the feature maps at resolutions other than 8×6 as inputs for each of the nine zero cross-attention blocks. Similarly, for the spatial encoder following the U-Net's encoder structure, we utilized the nine feature maps at resolutions other than 8×6 as the key and value inputs for the cross-attention layers.

Training & Inference Details. We train the model using an AdamW optimizer with a fixed learning rate of 1e-4 for 360k iterations, employing a batch size of 32. Then, we finetune the model with the attention total variation weight hyper-parameter $\lambda_{ATV} = 0.001$, using the same learning rate and batch size for 36K iterations. We train for about 100 hours using four NVIDIA A100 GPUs. For augmentation, we simultaneously applied horizontal flip (p=0.5) to both the clothing and the UNet's input condition, and independently applied Random Shift (limit=0.2, p=0.5) and Random Scale (limit=0.2, p=0.5) to both the clothing and UNet's input. We simultaneously applied HSV adjustments (limit=5, p=0.5) and contrast adjustments (limit=0.3, p=0.5) to both the clothing and $\mathbf{x}_0$. To prevent the issue of facial distortion, we finetune the decoder of the autoencoder separately on the training datasets VITON-HD [1] and DressCode [3]. In training the decoder, we use the AdamW optimizer with a learning rate of 5e-5 and a batch size of 32 for 10k iterations on each dataset. For inference, we employ the pseudo linear multi-step (PLMS) [2] sampler, with the number of sampling steps set to 50.

B. User Study Details.

In the user study, participants were asked to evaluate which of the two images, one generated by the baseline and the other by StableVITON, was superior in terms of 1) fidelity, 2) person attributes, 3) clothing identity, and 4) background quality (with respect to the cross-dataset setting). The questions for each criterion were as follows:

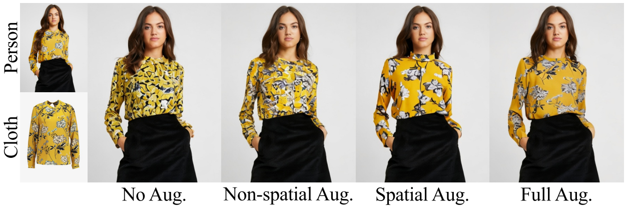

Figure 11. Visual ablation studies of the augmentations.

	SSIM	LPIPS	FID	KID
No Aug.	0.847	0.0969	9.35	1.33
Color Aug.	0.846	0.0986	9.33	1.26
Spatial Aug.	0.849	0.1025	9.12	1.17
Full Aug.	**0.850**	**0.0851**	**8.74**	**0.91**

Table 5. Ablation studies of the augmentation.

- Fidelity: Choose which image better exhibits resemblance to reality in aspects such as the human body and color harmony.
- Person Attributes: Choose which image better maintains features like skin tone, pose, and appearance from the input image.
- Clothing Identity: Choose which image better preserves characteristics such as the design, logo, and shape of the input clothing.
- Background Quality: Choose which image better maintains the background of the input image.

C. Additional Qualitative Results

In Fig. 12, we present the generation results on VITON-HD dataset, using the models trained on DressCode upper body dataset. GAN-based models (*i.e.*, HR-VITON and GP-VTON) show significant artifacts around the target person, as evidenced in quantitative results (see Table 2 in the main paper), leading to high FID and KID scores. In addition, diffusion-based models, while providing a plausible appearance, fail to preserve the details of the clothing.

D. Ablation Study of Augmentation

We divided the augmentations used in this paper into two categories: spatial (horizontal flip, random scale, and shift) and color (HSV and contrast), and trained the models separately. As shown in the Table 5 and Fig. 11, although spatial augmentation had a more significant impact than color augmentation, it was most effective to use both augmentations.

Thanks for the insightful comments, and we will incorporate the suggestions into the revised version.

E. StableVITON at High Resolution

To synthesize high-fidelity images, we have further trained StableVITON at the higher resolution of 1024×768. Instead of starting from scratch, we experimentally observed that progressively training StableVITON with 1024×768 resolution images leads to faster convergence. We conducted additional training for 85k iterations using the same training settings as StableVITON.

Qualitative Results. We present the results generated by StableVITON trained on VITON-HD dataset at 1024×768 resolution for images in VITON-HD, DressCode, SHHQ-1.0, and web-crawled datasets, in Fig. 14, Fig. 15, Fig. 16, and Fig. 17, respectively. We observe that there is a clearer preservation of facial or clothing details at 1024×768 resolution.

F. Limitations & Discussion

Even when fine-tuning the decoder of the autoencoder on the virtual try-on dataset, it remains challenging to preserve very fine details of the face or clothing. As a result, as shown in Fig. 5 of the main paper, subtle variations in facial features, such as the eyes, can be observed. However, we effectively address these issues related to fine details by increasing the model's resolution, as demonstrated in Fig. 14.

In our experiments, we observed that our model fails to preserve objects occluding the person or accessories such as bracelets and watches attached to the target person. This issue arises from the model's inability to incorporate additional information, apart from clothing, during the sampling process to fill the masked regions of the agnostic map. We leave such preservation issues as future work.

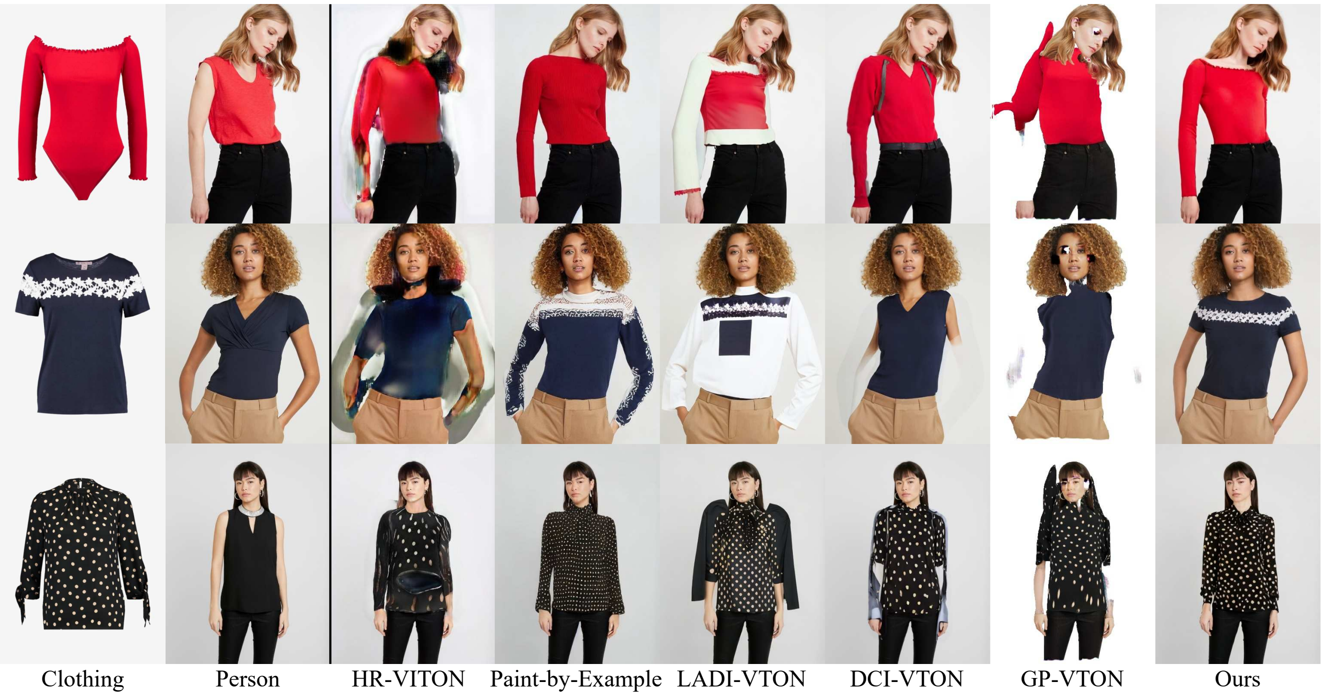

Figure 12. Qualitative comparison with baselines in a cross dataset setting (DressCode / VITON-HD). Best viewed when zoomed in.

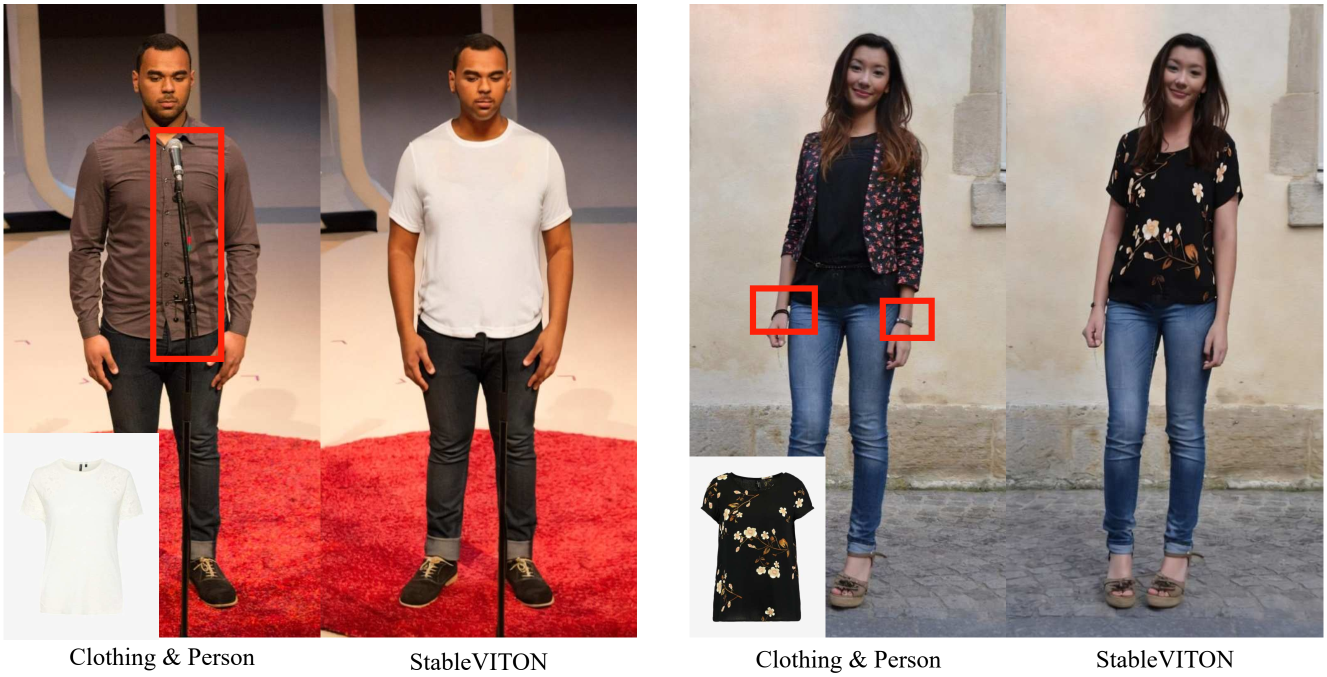

Figure 13. Limitations of StableVITON. StableVITON fails to preserve objects occluding the person or accessories such as bracelets.

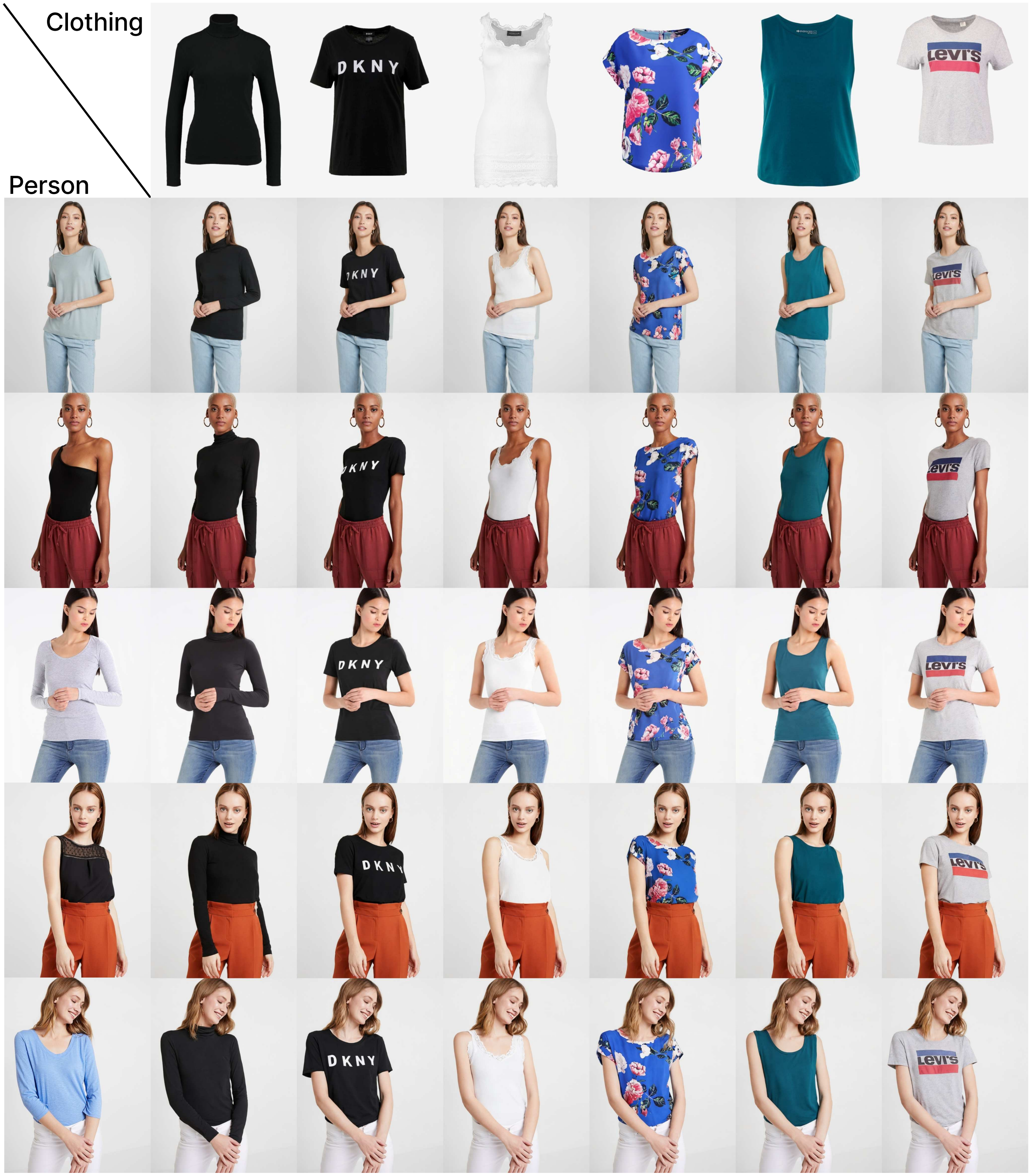

Figure 14. The generation results for the VITON-HD test dataset by StableVITON trained on VITON-HD dataset at 1024 × 768 resolution. Best viewed when zoomed in.

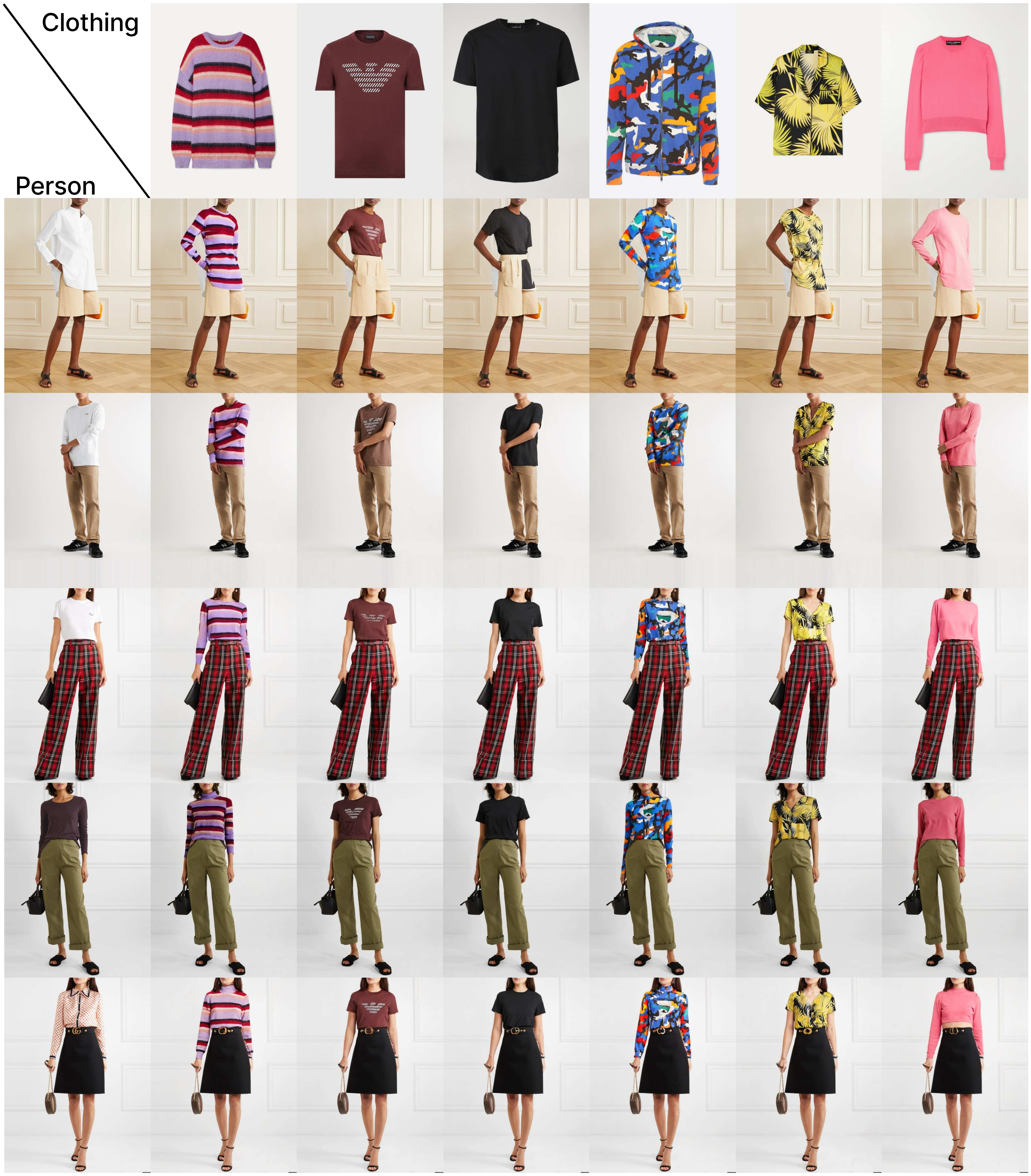

Figure 15. The generation results for the DressCode dataset by StableVITON trained on VITON-HD dataset at 1024 × 768 resolution. Best viewed when zoomed in.

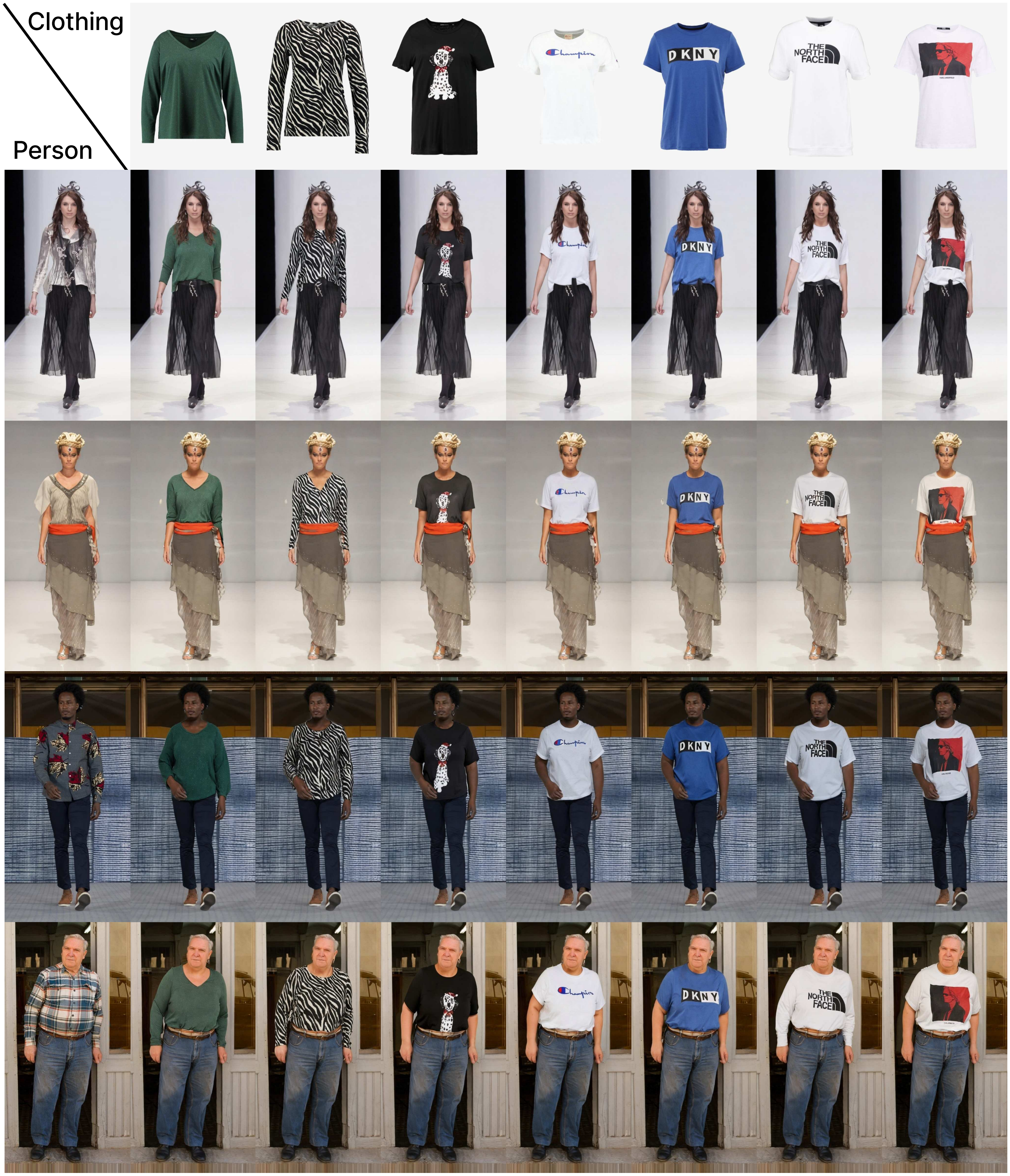

Figure 16. The generation results for the SHHQ-1.0 dataset by StableVITON trained on VITON-HD dataset at 1024 × 768 resolution. Best viewed when zoomed in.

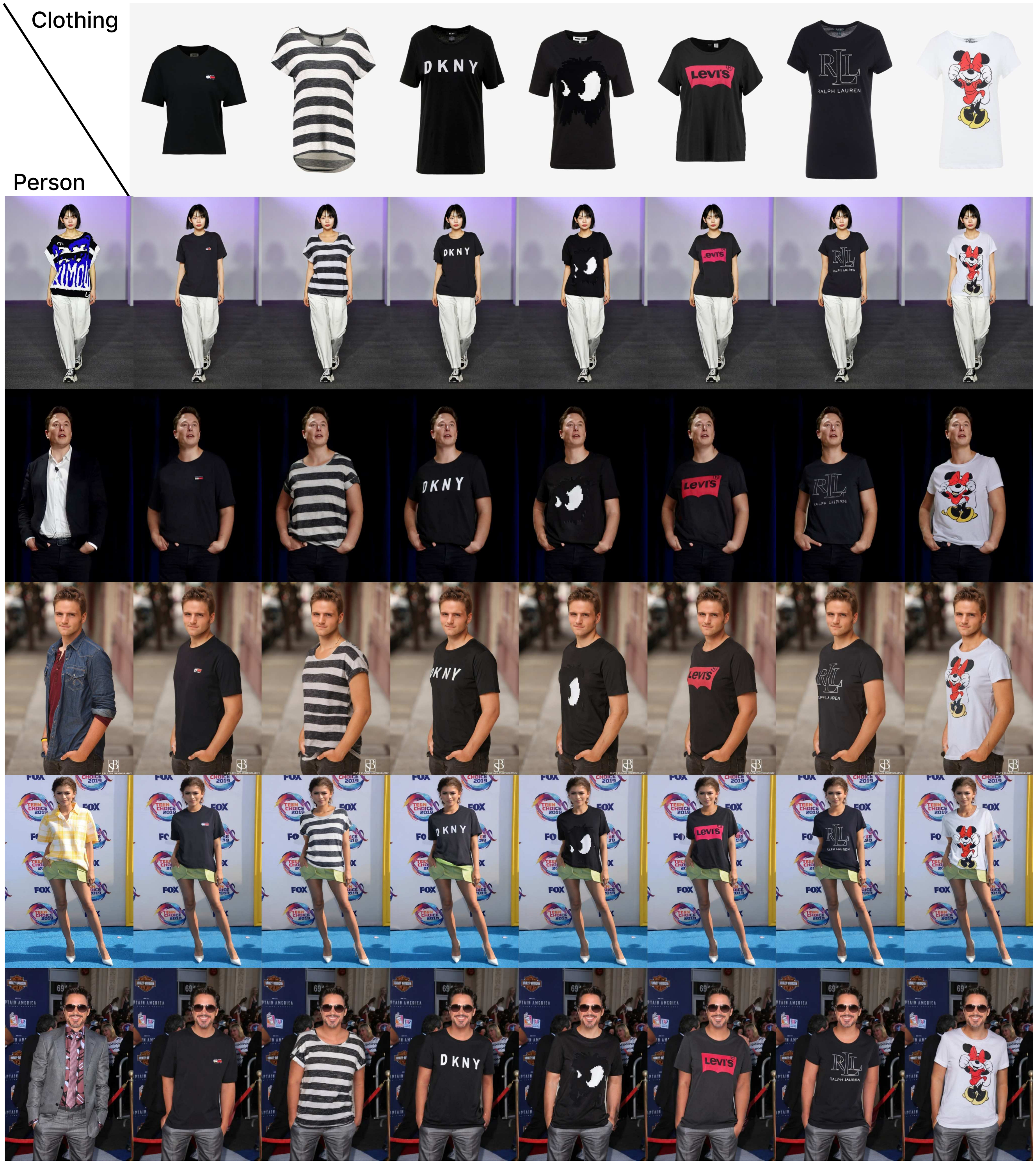

Figure 17. The generation results for the web-crawled images by StableVITON trained on VITON-HD dataset at 1024 × 768 resolution. Best viewed when zoomed in.

References

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, pages 14131–14140, 2021. 1

[2] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 1

[3] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *ECCV*, pages 2231–2235, 2022. 1

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[5] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 1