Supplemetary of Enhancing Virtual Try-On with Synthetic Pairs and Error-Aware Noise Scheduling

Parameter	Value
Batch Ratio of Synthetic Data	15%
Batch Size	32
Image Size	512x512
#Model Parameters	102.6M
Learning Rate	10^{-4}
#Training Iterations	200K
#Finetuning Iterations	100K

Table 6. Implementation details of EARSB+H2G-UH/FH.

A. Implementations Details

For generating the initial image x_1 in our EARSB training, we employ three try-on GAN models: HR-VTON [5] and SD-VTON [8] and GP-VTON [9]. All human images are processed to maintain their aspect ratio, with the longer side resized to 512 pixels and the shorter side padded with white pixels to reach 512. During training, images undergo random shifting and flipping with a 0.2 probability. The weakly-supervised classifier is trained for 100K iterations with a batch size of 8, while the human-to-garment GAN is trained for 90K iterations with a batch size of 16. As shown in Tab. 6, EARSB+H2G-UH/FH is trained for 300K iterations with a batch size of 32, incorporating 15% synthetic pairs in each batch. The first 200K iterations are trained on $t \in [0, 1]$ while the following 100k iterations are finetuned on $t \in [0, 0.5)$ and $t \in [0.5, 1]$ respectively following [1]. All models utilize the AdamW optimizer with a learning rate of 10^{-4}.

For inference, we select the GAN model that demonstrates better performance on each dataset to generate the initial image. Specifically, we employ GP-VTON [9] for VITON-HD and SD-VTON [8] for DressCode-Upper. During the sampling process, the guidance score in Eq. (10) is scaled by a factor of 6 and clamped to the range $[-0.3, 0.3]$.

B. UNet Architecture

EARSB UNet. The UNet architecture in EARSB consists of residual blocks and garment warping modules. It processes the concatenation of the error map M, pose representation P, and noisy image x_t to predict the noise distribution ϵ_θ^r at time t. The UNet encoder has 21 residual blocks, with the number of channels doubling every three blocks to a maximum of 256. Similarly, the garment encoder has 21 residual blocks but reaches a maximum of 128 channels. The decoder mirrors the encoder's structure, with extra garment warping modules. As shown in Fig. 8, each of the first 15 residual blocks in the UNet decoder is followed by a convolutional warping module. These modules concatenate encoded garment features and UNet-decoded features to predict a flow-like map for spatially warping the encoded garment features. The warped features are then injected into the subsequent decoder layer via input concatenation. Following [7], all residual blocks and flow-learning modules

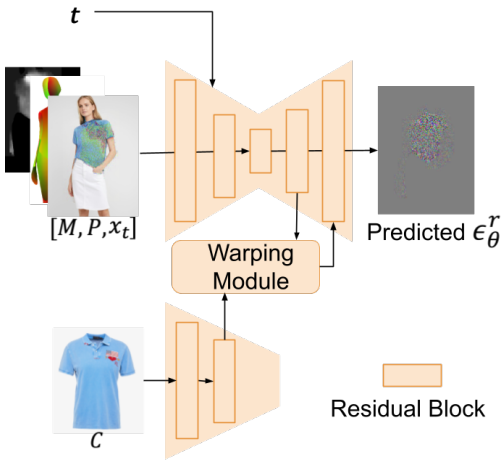

Figure 8. Architecture of our UNet in EARSB.

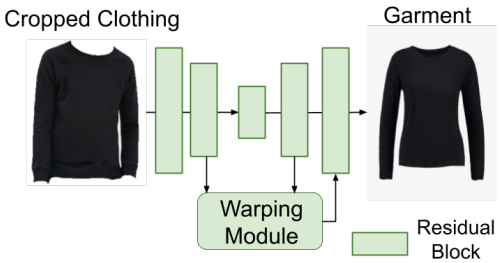

Figure 9. Architecture of our UNet in the human-to-garment model.

Figure 10. Results on different time steps. Our error map focuses on low-quality regions and maintains the quality of the sufficiently good regions.

incorporate timestep embeddings to renormalize latent features.

Human-to-Garment UNet. Our human-to-garment UNet architecture is adapted from the model proposed in [3]. As illustrated in Fig. 9, it shares similarities with the UNet in EARSB, but with two key distinctions: a) It is not timestep-dependent and takes cropped clothing as input to generate its product-view image. b) The garment warping module utilizes the i_{th} clothing features from both the encoder and decoder to learn a flow-like map, rather than using encoded features from the human.

C. Visualizing Error Maps

Our EARSB focuses on fixing specific errors and therefore can save the sampling cost when initial predictions are sufficiently good. For example, in the first row of Fig. 10, the error map highlights the graphics and text in the initial image. This low-quality part is being refined progressively as the number of sampling steps increases from 5 to 100. At the same time, other parts that our weakly-supervised classifier believes to be sufficiently good, which are mostly the solid-color areas, are kept well regardless of the number of sam-

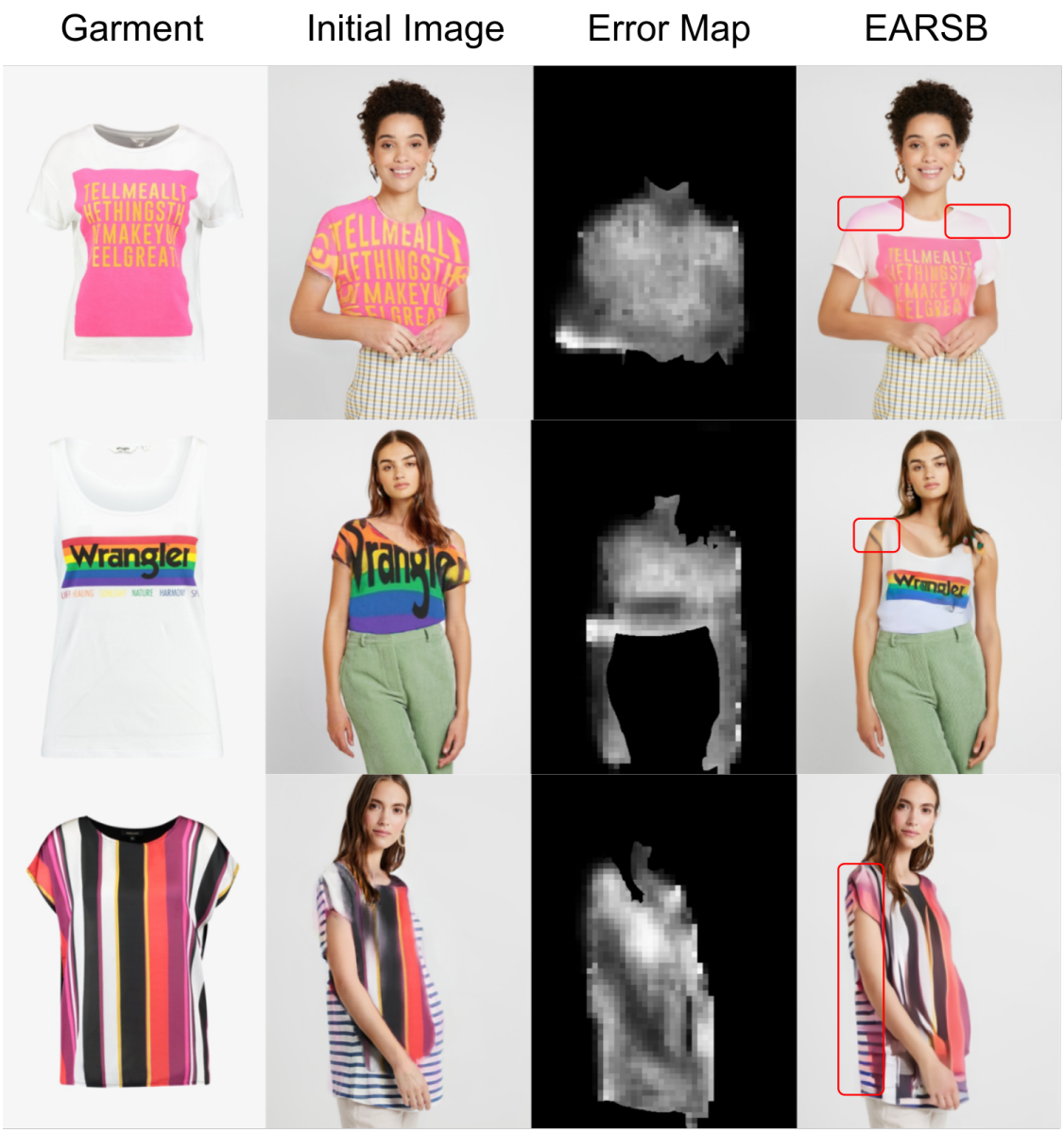

Figure 11. Failure cases on VITON-HD where the initial image has a poor-quality.

pling steps. Therefore, for an initial image whose error map has almost zero values, we can choose to use fewer steps in sampling. On the contrary, for an initial image whose error map has high confidence, we should assign more sampling steps to it to improve the image quality.

D. Ablations on the Quality of the Initial Image

In Tab. 7 we include the FID results of using different try-on GAN models to generate the initial image under the unpaired setting. Baseline means the GAN baseline. We can draw three conclusions from the results: a) our EARSB can refine the GAN-generated image over the GAN baseline; b) the quality of the initial image x_1 is positively correlated with the quality of the sampled $\hat{x}_0$; c) our model achieves higher gains over CAT-DM, which also tries to refine the

	HR-VTON [5]	SD-VTON [8]	GP-VTON [9]
Baseline	10.75	9.05	8.61
CAT-DM [10]	10.03	8.76	8.55
EARSB	**9.11**	**8.69**	**8.42**

Table 7. FID scores of using different try-on GAN models to generate the initial image under the unpaired setting.

Figure 12. Visualization of the generated images in WVTON.

	FID↓	KID↓
Stable-VTON [4]	131.76	2.10
EARSB(SD)	127.15	1.67
EARSB(SD) +H2G-UH	**120.29**	**1.18**

Table 8. Results on out-of-domain test set WVTON under the unpaired setting. All image background is removed for evaluation.

GAN-generated image but without error-aware noise schedule.

E. Results on In-the-Wild Dataset

We ran our data-augmented EARSB on the Out-of-Domain test set WVTON [6] under the unpaired setting and removed the image background for evaluation. In Tab. 8, we observe a 7 point gain in FID, showing its good generalization ability. Fig. 12 also shows that EARSB(SD)+H2G-UH better recovers the clothing patterns.

F. Limitations

While our human-to-garment model can effectively generate synthetic paired data for try-on training augmentation, it has some imperfections. The overall quality of synthetic garments is regulated by our filtering criteria (Sec. 3.2), yet minor texture deformations occasionally occur. For instance, in Fig. 13, the second pair of the first row shows a misaligned shirt placket in the synthetic garment. This limitation stems partly from the fact that our model is trained in the image domain which lacks 3D information. A potential solution is to utilize DensePose representations extracted from the garment as in [2].

A key constraint of our EARSB is its refinement-based nature, which makes the generated image dependent on the initial image. We assume that the initial image from a try-on GAN model is of reasonable quality, requiring only partial refinement. Consequently, if the initial image is of very poor quality, our refinement process cannot completely erase and regenerate an entirely new, unrelated image. Fig. 11 illustrates this limitation: in the first row, the initial image severely mismatches the white shirt with pink graphics. With EARSB refinement, while the shirt is correctly rewarped, color residuals from the initial image persist around the shoulder area.

	FID	KID	SSIM	LPIPS
VITON-HD	14.81	0.42	0.849	0.229
DressCode-Upper	18.92	0.59	0.832	0.257

Table 9. Human-to-Garment results under 1024x1024 image resolution.

G. Additional Visualizations

Figures 13 and 14 showcase exemplars from our synthesized datasets H2G-UH and H2G-FH, respectively. We also report quantitative results in Table 9 to evaluate our human-to-garment model on VITON-HD and DressCode-Upper. The generated garment images in Figures 13 and 14 closely mimic the product view of the clothing items, accurately capturing both the shape and texture of the original garments worn by the individuals. This approach to creating synthetic training data for the virtual try-on task is both cost-effective and data-efficient, highlighting the benefits of our proposed human-to-garment model.

Figures 15 and 16 give visualized results of the proposed EARSB and EARSB+H2G-UH. In contrast to previous approaches, EARSB specifically targets and enhances low-quality regions in GAN-generated images, which typically correspond to texture-rich areas. This targeted improvement is evident in the last row of Fig. 15, where EARSB more accurately reconstructs text *freinds*, and in the third row, where it successfully generates four side buttons. Furthermore, the incorporation of our synthetic dataset H2G-UH with EARSB leads to even more refined details in the generated images, demonstrating the synergistic effect of our combined approach.

H. Ethics

We acknowledge several potential ethical considerations of our work on virtual try-on:

- Bias and representation: We strive for diversity in our training data to ensure the model performs equitably across different body types, skin tones, and ethnicities. However, biases may still exist, and further work is needed to assess and mitigate these.
- Misuse potential: While intended for benign purposes, this technology could potentially be misused to create misleading or non-consensual images. We strongly condemn such uses and will explore safeguards against misuse in future work.

We believe the potential benefits of this technology outweigh the risks, but we remain vigilant about these ethical considerations and are committed to addressing them as our research progresses.

Figure 13. Visualized examples of the (human, synthetic garment) pairs on our proposed H2G-UH.

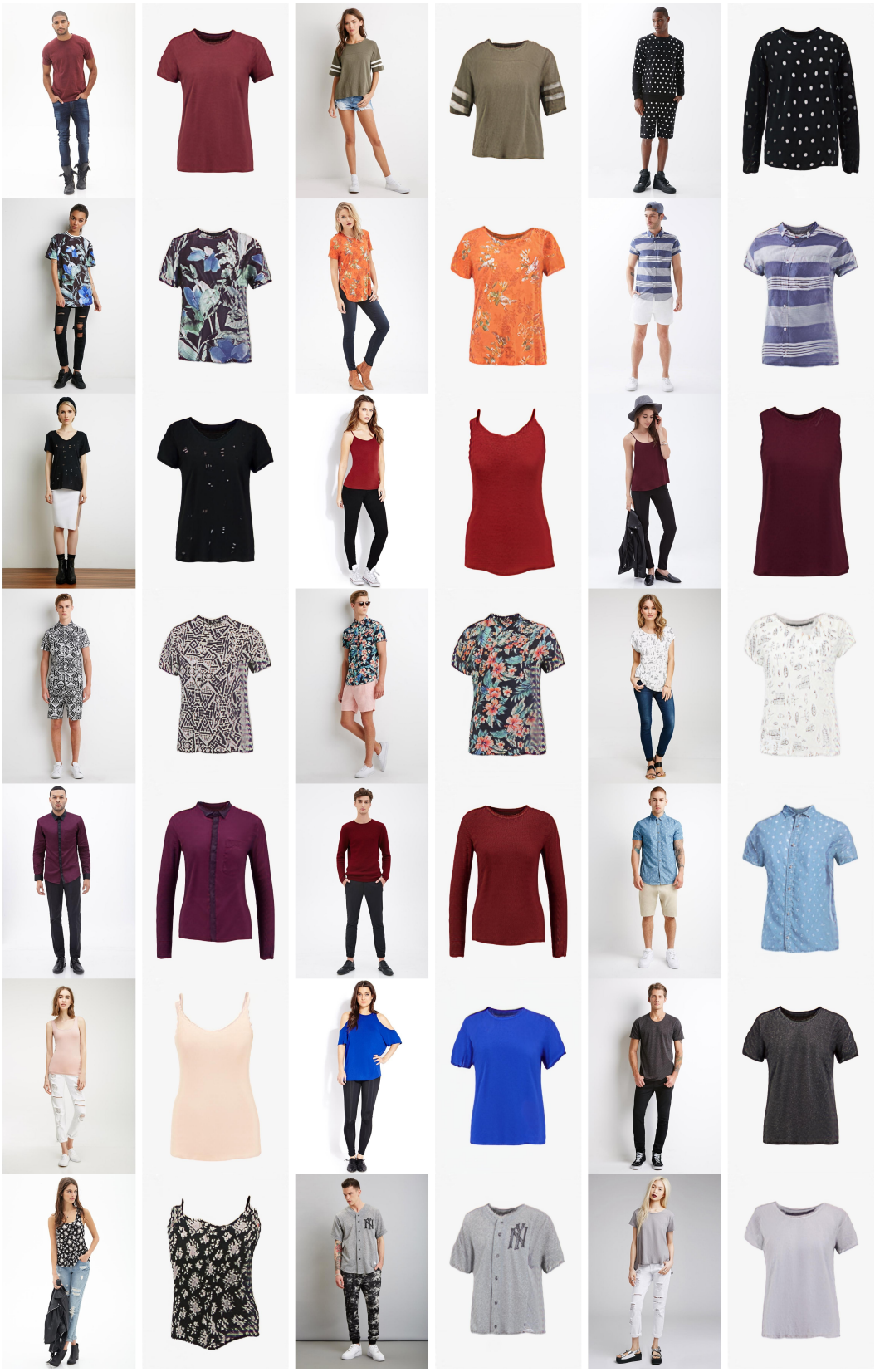

Figure 14. Visualized examples of the (human, synthetic garment) pairs on our proposed H2G-FH.

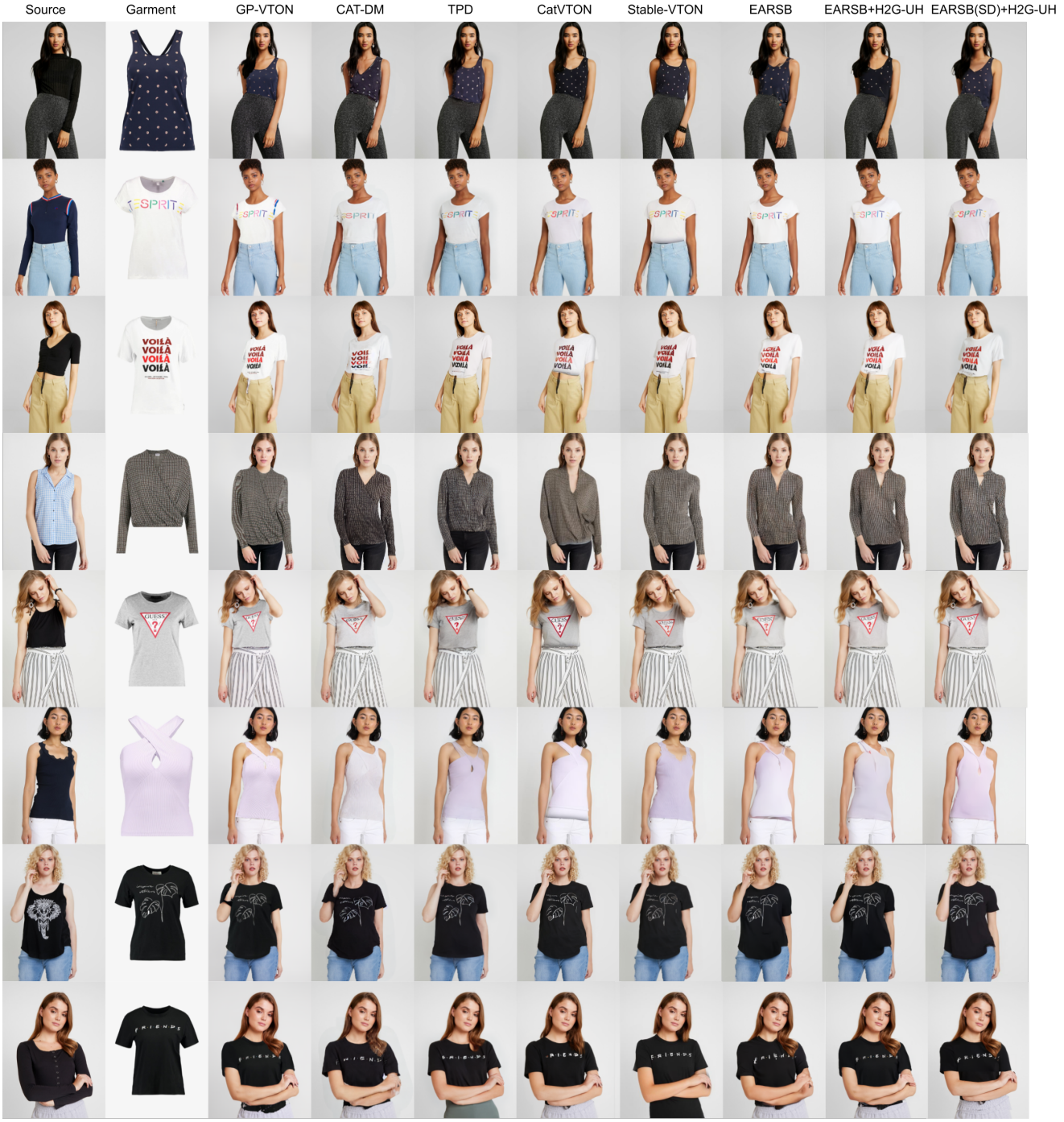

Figure 15. Visualized examples on VITON-HD. Our EARSB and EARSB+H2G-UH better recovers the intricate textures in the garment.

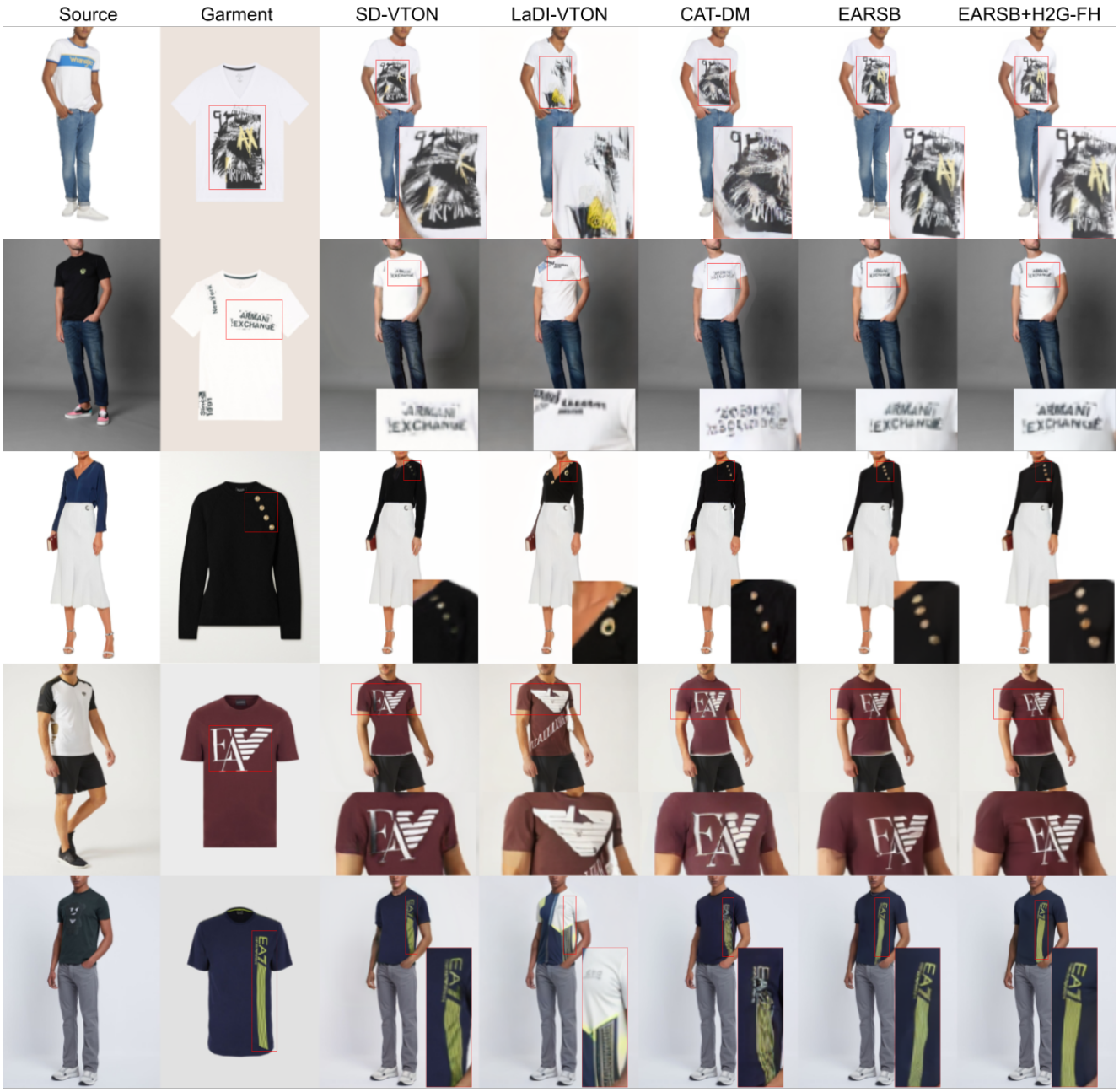

Figure 16. Visualized examples on DressCode-Upper. Our EARSB and EARSB+H2G-UH better reconstructs the texts and graphics in the garment.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[2] Aiyu Cui, Sen He, Tao Xiang, and Antoine Toisoul. Learning garment densepose for robust warping in virtual try-on. *arXiv preprint arXiv:2303.17688*, 2023.

[3] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *CVPR*, 2019.

[4] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. StableVITON: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, 2024.

[5] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, 2022.

[6] Nannan Li, Qing Liu, Krishna Kumar Singh, Yilin Wang, Jianming Zhang, Bryan A Plummer, and Zhe Lin. UniHuman: A unified model for editing human images in the wild. In *CVPR*, 2024.

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

[8] Sang-Heon Shim, Jiwoo Chung, and Jae-Pil Heo. Towards squeezing-averse virtual try-on via sequential deformation. In *AAAI*, 2024.

[9] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. GP-VTON: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *CVPR*, 2023.

[10] Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *CVPR*, 2024.