TryOnDiffusion: A Tale of Two UNets
Supplementary Material

Luyang Zhu[1,2*] Dawei Yang[2] Tyler Zhu[2] Fitsum Reda[2] William Chan[2]
Chitwan Saharia[2] Mohammad Norouzi[2] Ira Kemelmacher-Shlizerman[1,2]
[1]University of Washington [2]Google Research

1. Implementation Details

1.1. Parallel-UNet

Fig. 1 provides the architecture of 256×256 Parallel-UNet. Compared to the 128×128 version, 256×256 Parallel-UNet makes the following changes: 1) In addition to the try-on conditional inputs $\mathbf{c}_{\text{tryon}}$, the 256×256 Parallel-UNet takes as input the try-on result I_{tr}^{128}, which is first bilinearly upsampled to 256×256, and then concatenated to the noisy image $\mathbf{z}_t$; 2) the self attention and cross attention modules only happen at 16×16 resolution; 3) extra UNet blocks at 256×256 resolution are used; 4) the repeated times of UNet blocks are different as indicated by the Figures.

For both 128×128 and 256×256 Parallel-UNet, normalization layers are parametrized as Group Normalization [9]. The number of group is set to $\min(32, \lfloor\frac{C}{4}\rfloor)$, where C is the number of channels for input features. The non-linear activation is set to swish [5] across the whole model. The residual blocks used in each scale have a main pathway of GroupNorm→swish→conv→GroupNorm→swish→conv. The input to the residual block is processed by a separate convolution layer and added to the output of the main pathway as the skip connection. The number of feature channels for UNet blocks in 128×128 Parallel-UNet is set to $128, 256, 512, 1024$ for resolution $128, 64, 32, 16$ respectively. The number of feature channels for UNet blocks in 256×256 Parallel-UNet is set to $128, 128, 256, 512, 1024$ for resolution $256, 128, 64, 32, 16$ respectively. The positional encodings of diffusion timstep t and noise augmentation levels $\mathbf{t}_{\text{na}}$ are not shown in the figures for cleaner visualization. They are used for FiLM [4] as described in Sec. 3.2 of the main paper. The 128×128 Parallel-UNet has 1.13B parameters in total while the 256×256 Parallel-UNet has 1.06B parameters.

1.2. Training and Inference

TryOnDiffusion was implemented in JAX [2]. All three diffusion models are trained on 32 TPU-v4 chips for 500K iterations (around 3 days for each diffusion model). After trained, we run the inference of the whole pipeline on 4 TPU-v4 chips with batch size 4, which takes around 18 seconds for one batch.

2. Additional Results

In Fig. 2 and 3, we provide qualitative comparison to state-of-the-art methods on challenging cases. We select input pairs from our 6K testing dataset with heavy occlusions and extreme body pose and shape differences. We can see that our method can generate more realistic results compared to baselines. In Fig. 4 and 5, we provide qualitative comparison to state-of-the-art methods on simple cases. We select input pairs from our 6K test dataset with minimum garment warp and simple texture pattern. Baseline methods perform better for simple cases than for challenging cases. However, our method is still better at garment detail preservation and blending (of person and garment). In Fig. 6, we provide more qualitative results on the VITON-HD unpaired testing dataset.

For fair comparison, we run a new user study to compare SDAFN [1] vs our method at SDAFN's 256 × 256 resolution. To generate a 256 × 256 image with our method, we only run inference on the first two stages of our cascaded diffusion models and ignore the 256×256→1024×1024 SR diffusion. Table 1 shows results consistent with the user study reported in the paper. We also compare to HR-VITON [7] using their released checkpoints. Note that original HR-VTION is trained on frontal garment images, so we select input garments satisfying this constraint to avoid unfair comparison. Fig. 9 shows that our method is still better than HR-VITON under its optimal cases using its released checkpoints.

Table 2 reports quantitative results for ablation studies. Fig. 7 visualizes more examples for the ablation study of combining warp and blend versus sequencing the tasks.

[1] Work done while author was an intern at Google.

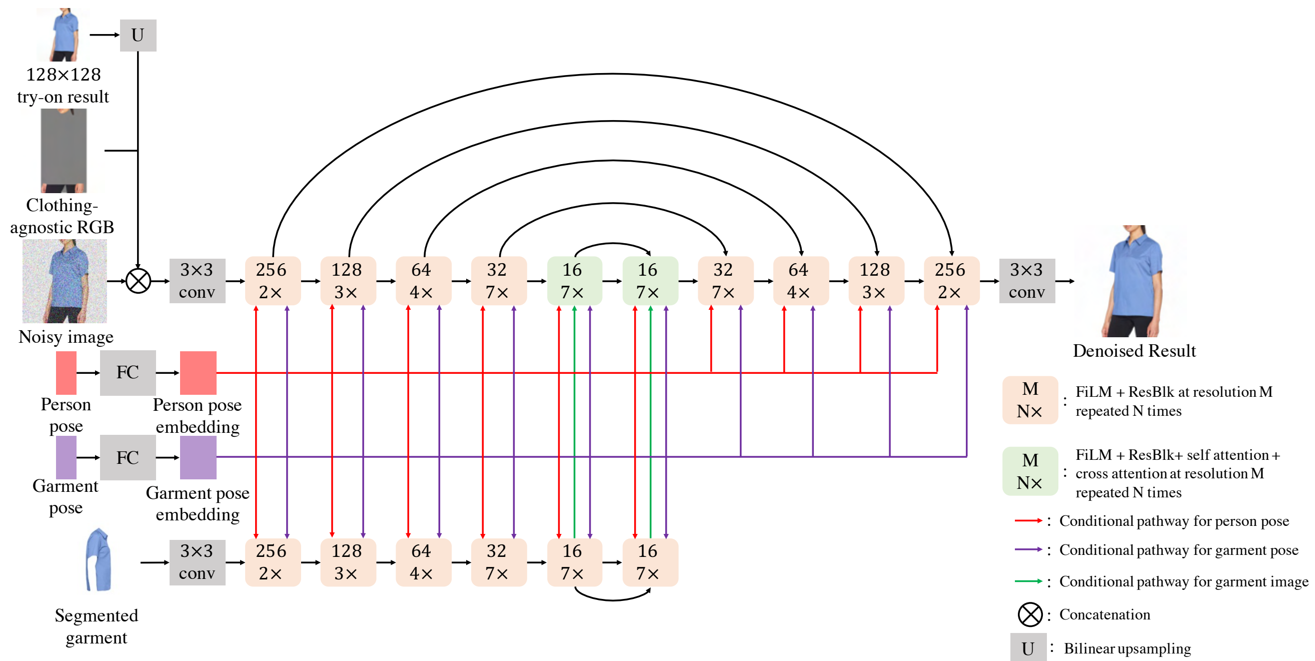

Figure 1. Architecture of 256×256 Parallel-UNet.

Fig. 8 provides more qualitative comparisons between concatenation and cross attention for implicit warping.

We further investigate the effect of the training dataset size. We retrained our method from scratch on 10K and 100K random pairs from our 4M set and report quantitative results (FID and KID) on two different test sets in Table 3. Fig. 10 also shows visual results for our models trained on different dataset sizes.

In Fig. 6 of the main paper, we provide failure cases due to erroneous garment segmentation and garment leaks in the clothing-agnostic RGB image. In Fig. 11, we provide more failure cases of our method. The main problem lies in the clothing-agnostic RGB image. Specifically, it removes part of the identity information from the target person, e.g., tattoos (row one), muscle structure (row two), fine hair on the skin (row two) and accessories (row three). To better visualize the difference in person identity, Fig. 12 provides try-on results on paired unseen test samples, where groundtruth is available.

Fig. 13 shows try-on results for a challenging case, where input person wearing garment with no folds, and input garment with folds. We can see that our method can generate realistic folds according to the person pose instead of copying folds from the garment input. Fig. 14 and 15 show TryOnDiffusion results on variety of people and garments for both men and women.

Finally, Fig. 16 to 21 provide zoom-in visualization for Fig. 1 of the main paper, demonstrating high quality results of our method.

	SDAFN [1]	Ours	Hard to tell
Random	5.24%	**77.83%**	16.93%
Challenging	3.96%	**93.99%**	2.05%

Table 1. User study comparing SDAFN [1] to our method at 256×256 resolution.

Test datasets	Ours		VITON-HD	
Methods	FID ↓	KID ↓	FID ↓	KID ↓
Ablation 1	15.691	7.956	25.093	12.360
Ablation 2	14.936	7.235	28.330	17.339
Ours	**13.447**	**6.964**	**23.352**	**10.838**

Table 2. Quantitative comparison for ablation studies. We compute FID and KID on our 6K test set and VITON-HD's unpaired test set. The KID is scaled by 1000 following [6].

Test datasets	Ours		VITON-HD	
Train set size	FID ↓	KID ↓	FID ↓	KID ↓
10K	16.287	8.975	25.040	11.419
100K	14.667	7.073	23.983	**10.732**
4M	**13.447**	**6.964**	**23.352**	10.838

Table 3. Quantitative results for the effects of the training set size. We compute FID and KID on our 6K test set and VITON-HD's unpaired test set. The KID is scaled by 1000 following [6].

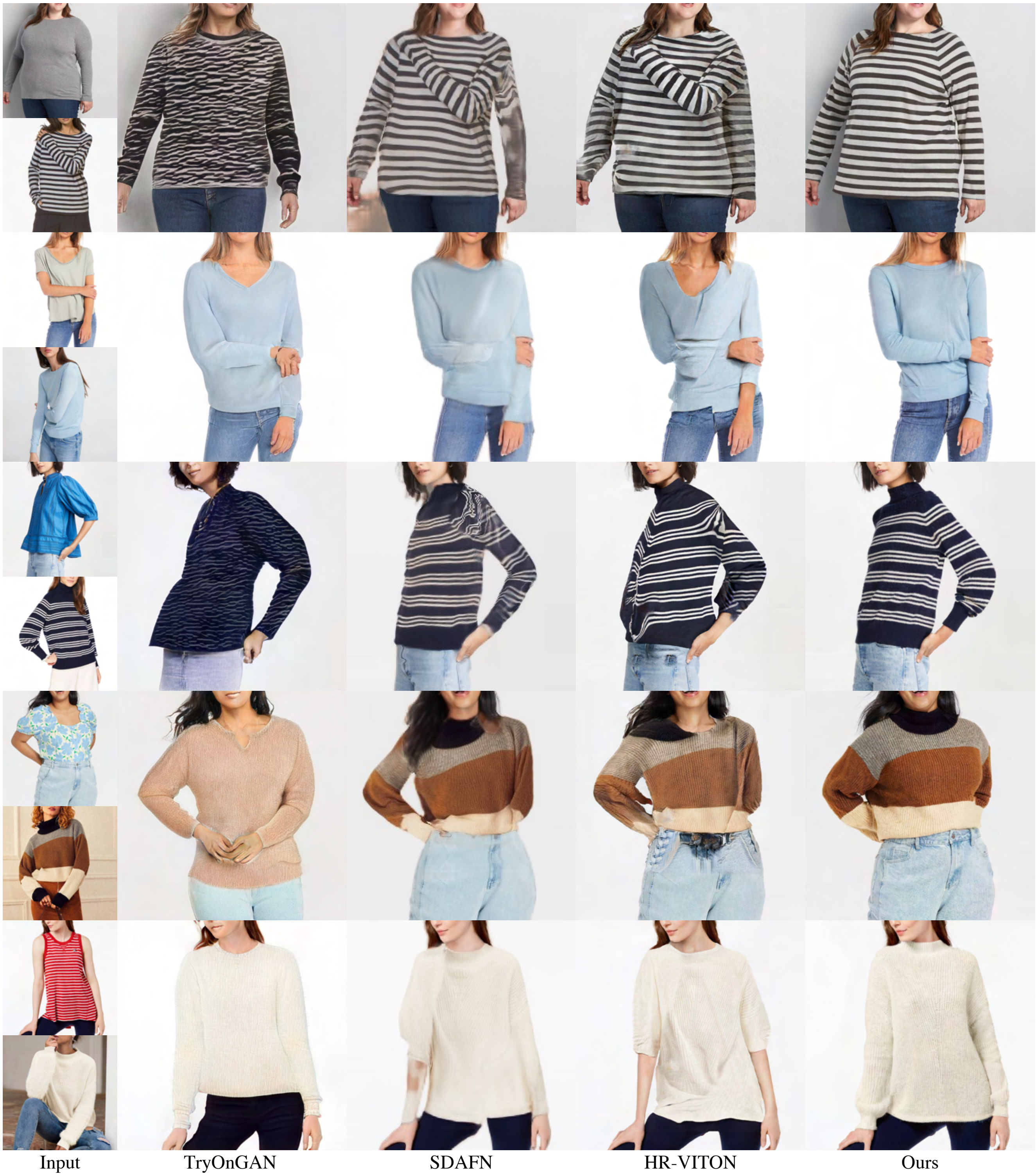

Figure 2. Comparison with TryOnGAN [8], SDAFN [1] and HR-VITON [7] on challenging cases for women. Compared to baselines, TryOnDiffusion can preserve garment details for heavy occlusions as well as extreme body pose and shape differences. Please zoom in to see details.

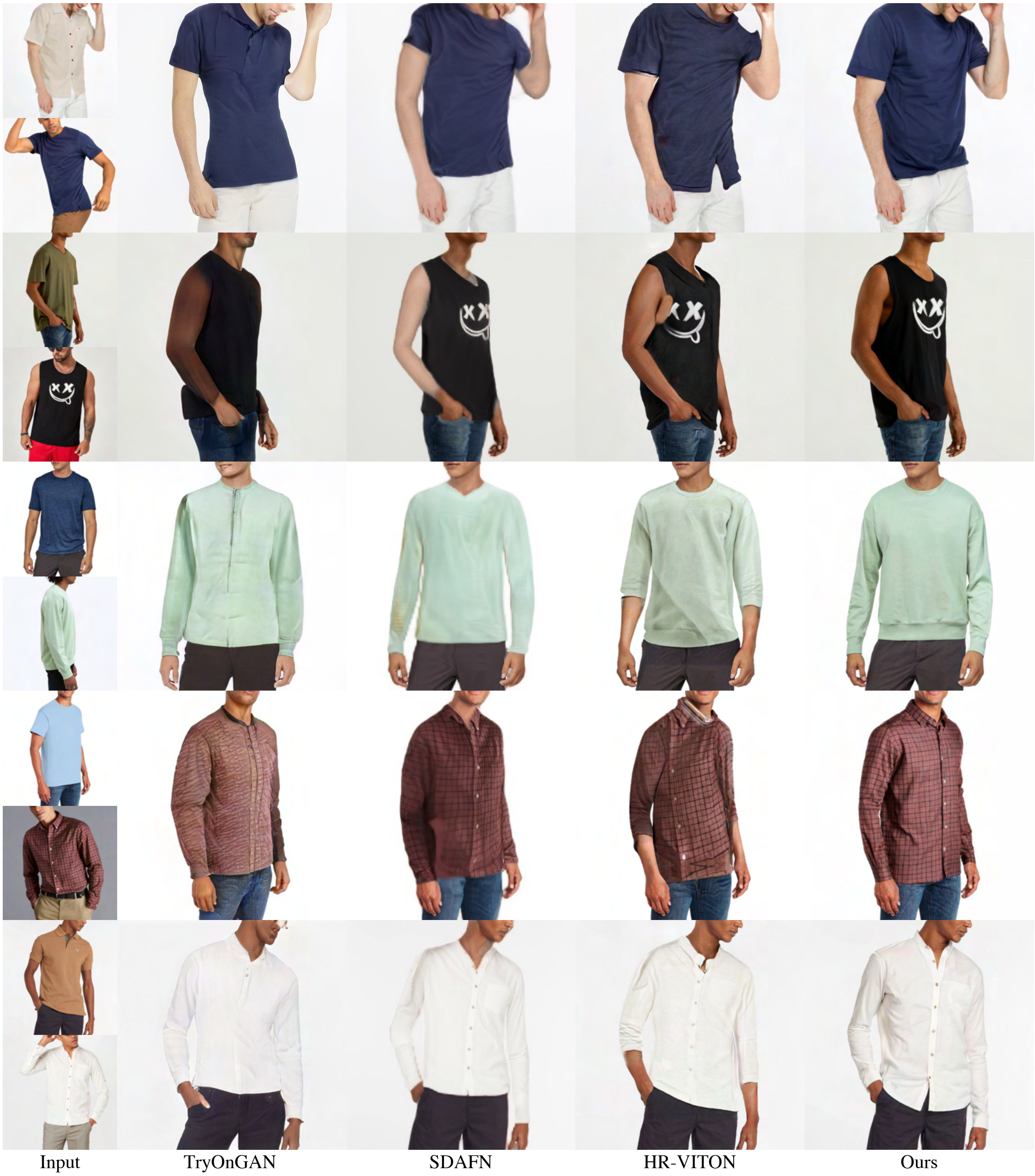

Figure 3. Comparison with TryOnGAN [8], SDAFN [1] and HR-VITON [7] on challenging cases for men. Compared to baselines, TryOnDiffusion can preserve garment details for heavy occlusions as well as extreme body pose and shape differences. Please zoom in to see details.

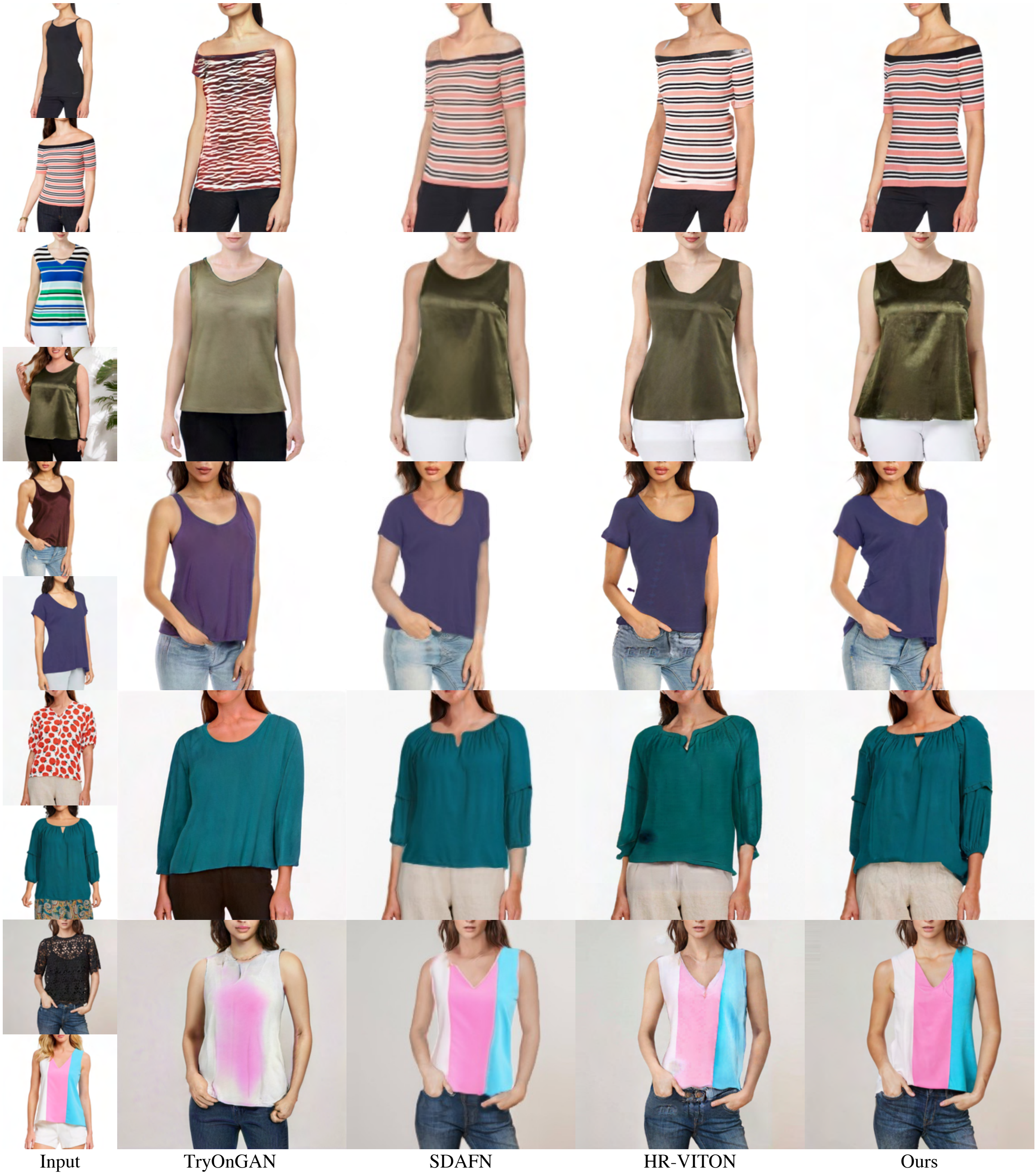

Figure 4. Comparison with TryOnGAN [8], SDAFN [1] and HR-VITON [7] on simple cases for women. We select input pairs with minimum garment warp and simple texture pattern. Baseline methods perform better for simple cases than for challenging cases. However, our method is still better at garment detail preservation and blending (of person and garment). Please zoom in to see details.

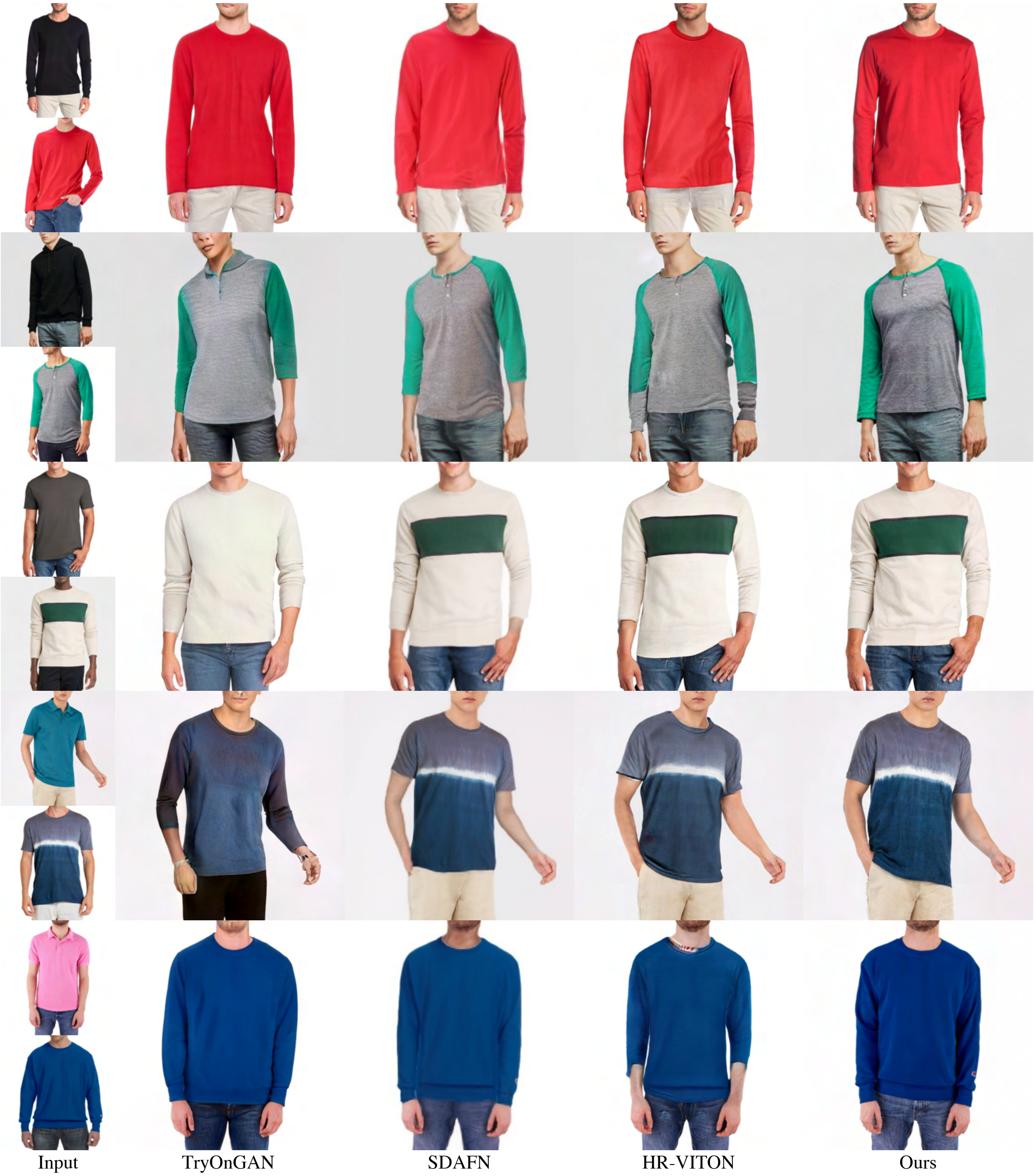

Figure 5. Comparison with TryOnGAN [8], SDAFN [1] and HR-VITON [7] on simple cases for men. We select input pairs with minimum garment warp and simple texture pattern. Baseline methods perform better for simple cases than for challenging cases. However, our method is still better at garment detail preservation and blending (of person and garment). Please zoom in to see details.

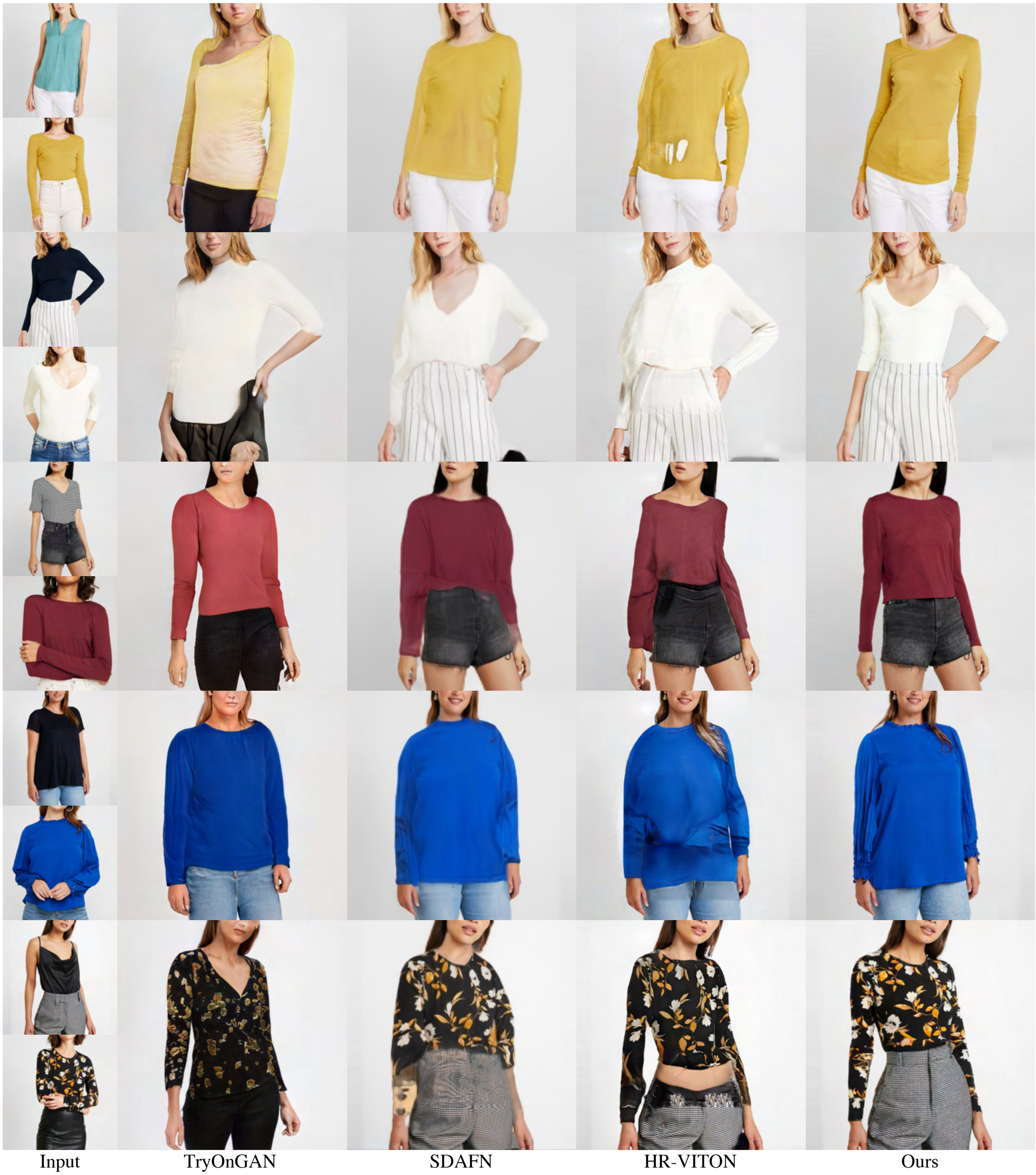

Figure 6. Comparison with state-of-the-art methods on VITON-HD unpaired testing dataset [3]. All methods were trained on the same 4M dataset and tested on VITON-HD. Please zoom in to see details

References

[1] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022. 1, 2, 3, 4, 5, 6

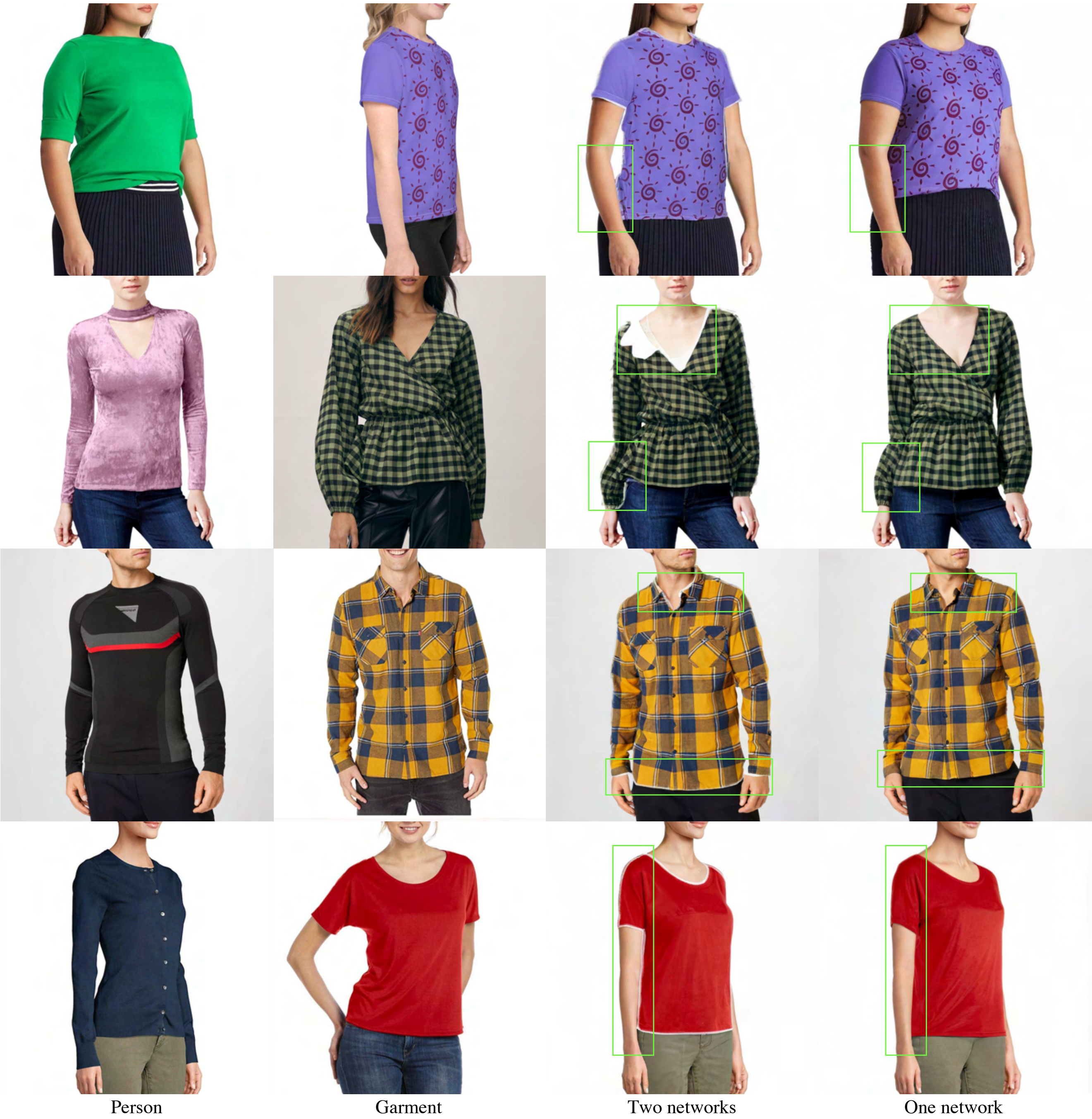

Figure 7. Combining warp and blend vs sequencing two tasks. Two networks (column 3) represent sequencing two tasks. One network (column 4) represents combining warp and blend. Green boxes highlight differences, please zoom in to see details.

[2] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 1

[3] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2021. 7

[4] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio.

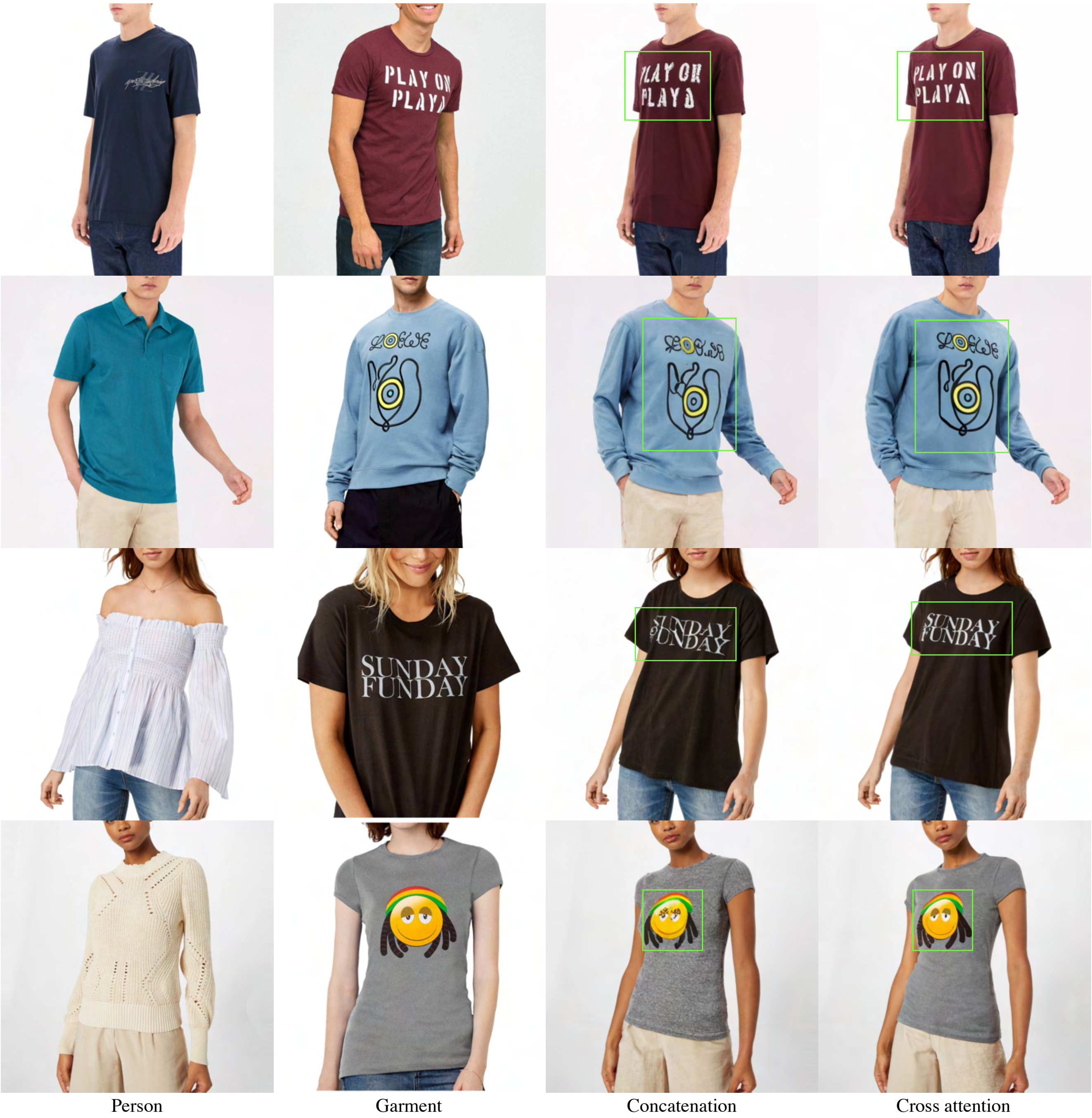

Figure 8. Cross attention vs concatenation for implicit warping. Green boxes highlight differences, please zoom in to see details.

Feature-wise transformations. *Distill*, 3(7):e11, 2018. 1

[5] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. 1

[6] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33:12104–12114, 2020. 2

[7] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 1, 3, 4, 5, 6

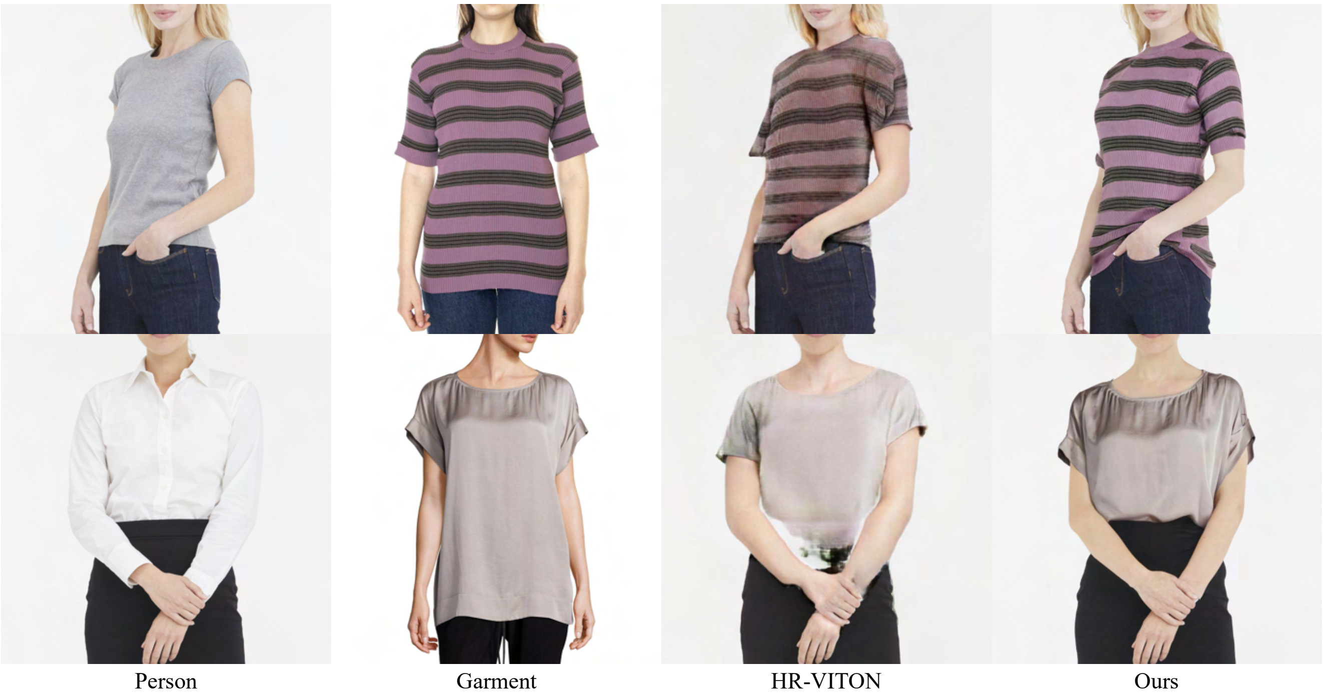

Figure 9. Comparison with HR-VITON released checkpoints for frontal garment (optimal for HR-VITON). Please zoom in to see details.

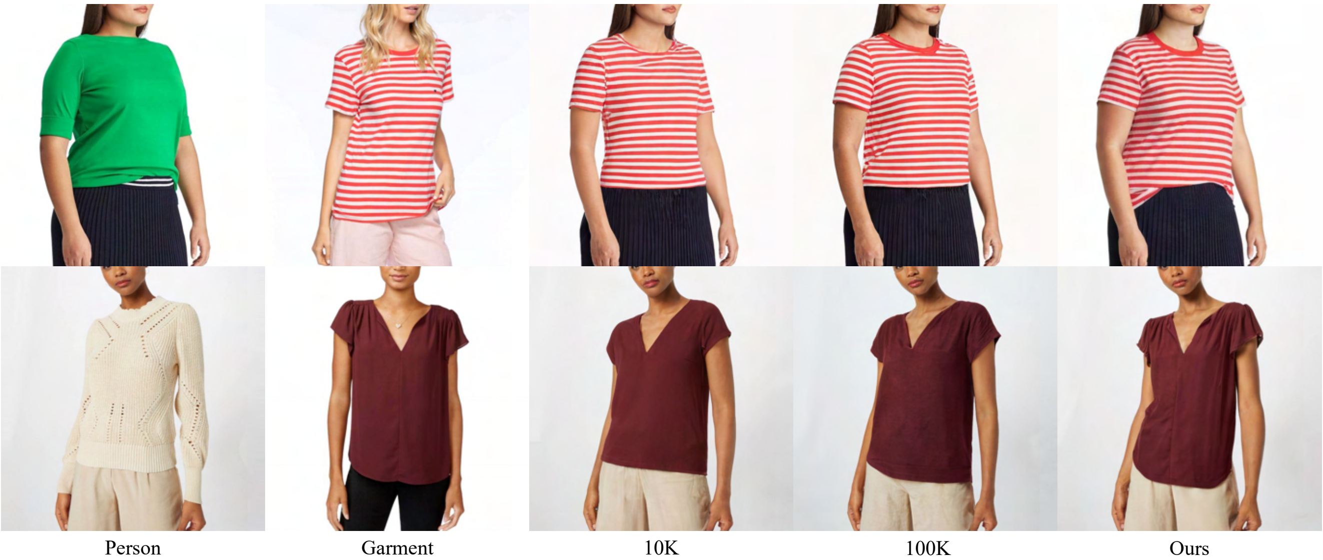

Figure 10. Quanlitative results for effects of the training set size. Please zoom in to see details.

[8] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021. 3, 4, 5, 6

[9] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

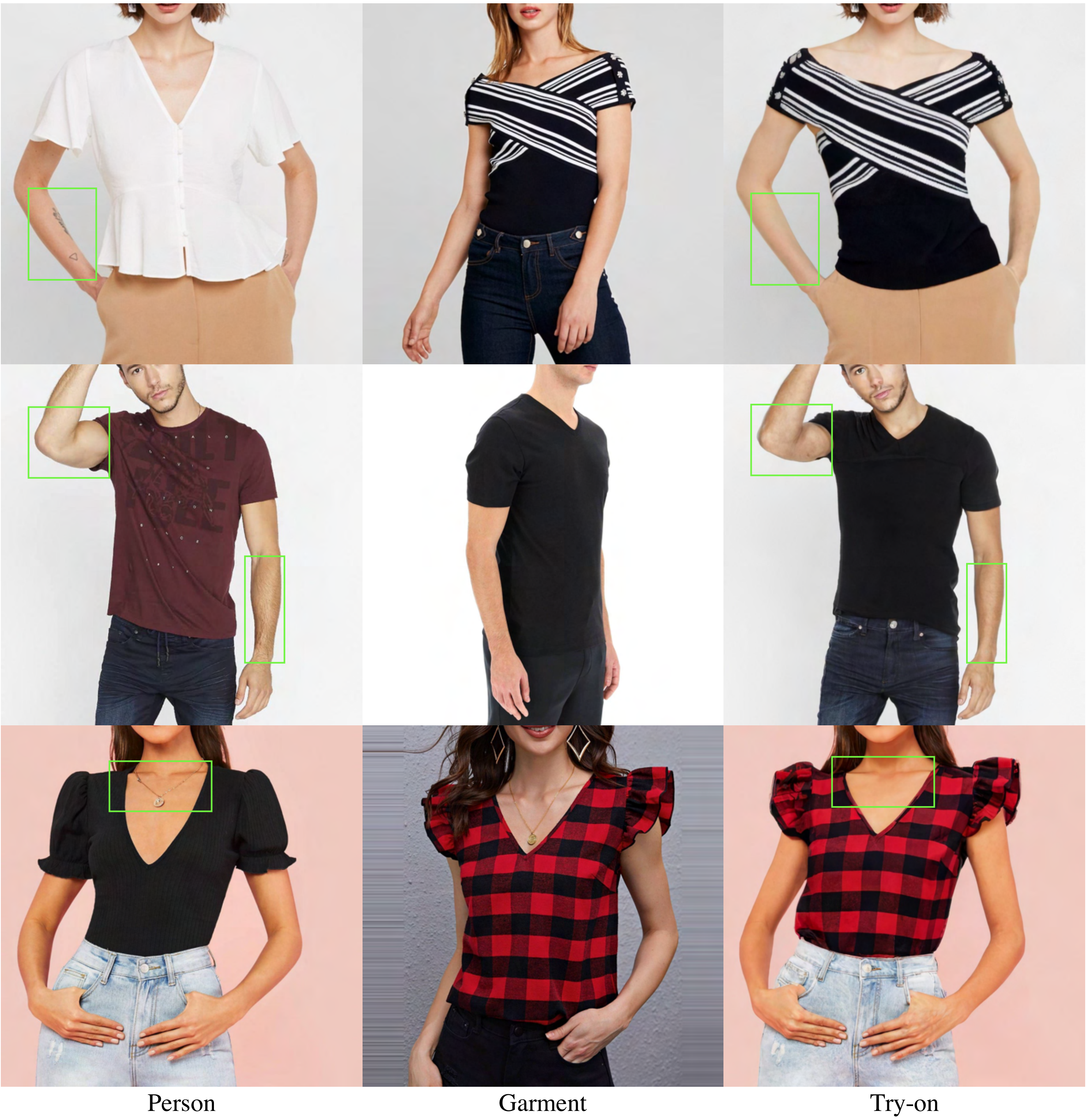

Figure 11. Failure cases. Clothing-agnostic RGB image removes part of the identity information from the target person, e.g., tattoos (row one), muscle structure (row two), fine hair on the skin (row two) and accessories (row three).

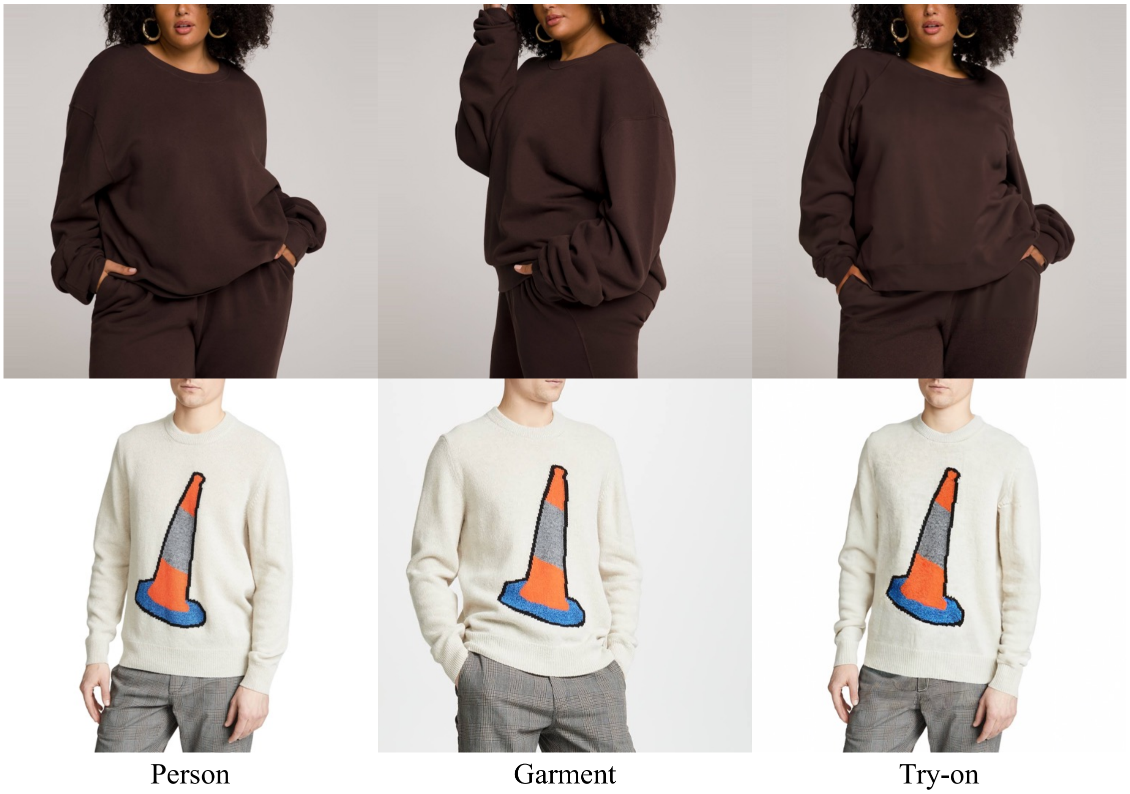

Figure 12. Qualitative results on paired unseen test samples. Please zoom in to see details.

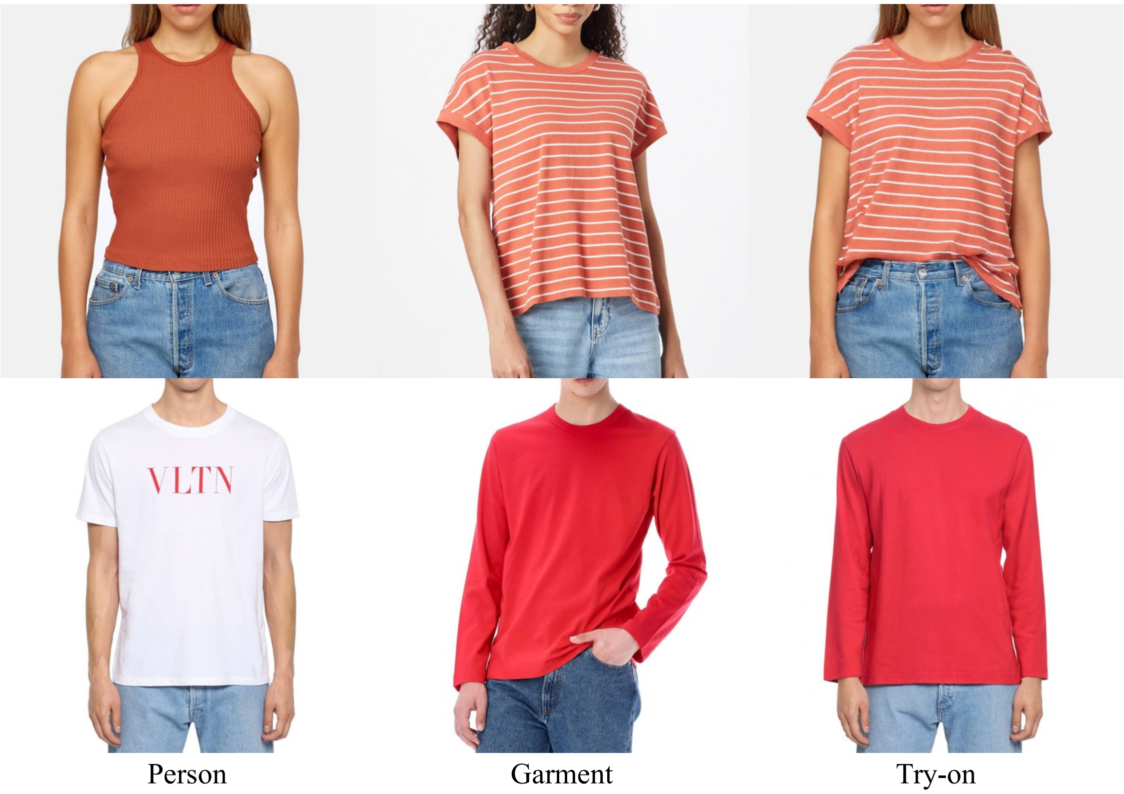

Figure 13. Try-on results for input person wearing garment with no folds, and input garment with folds.

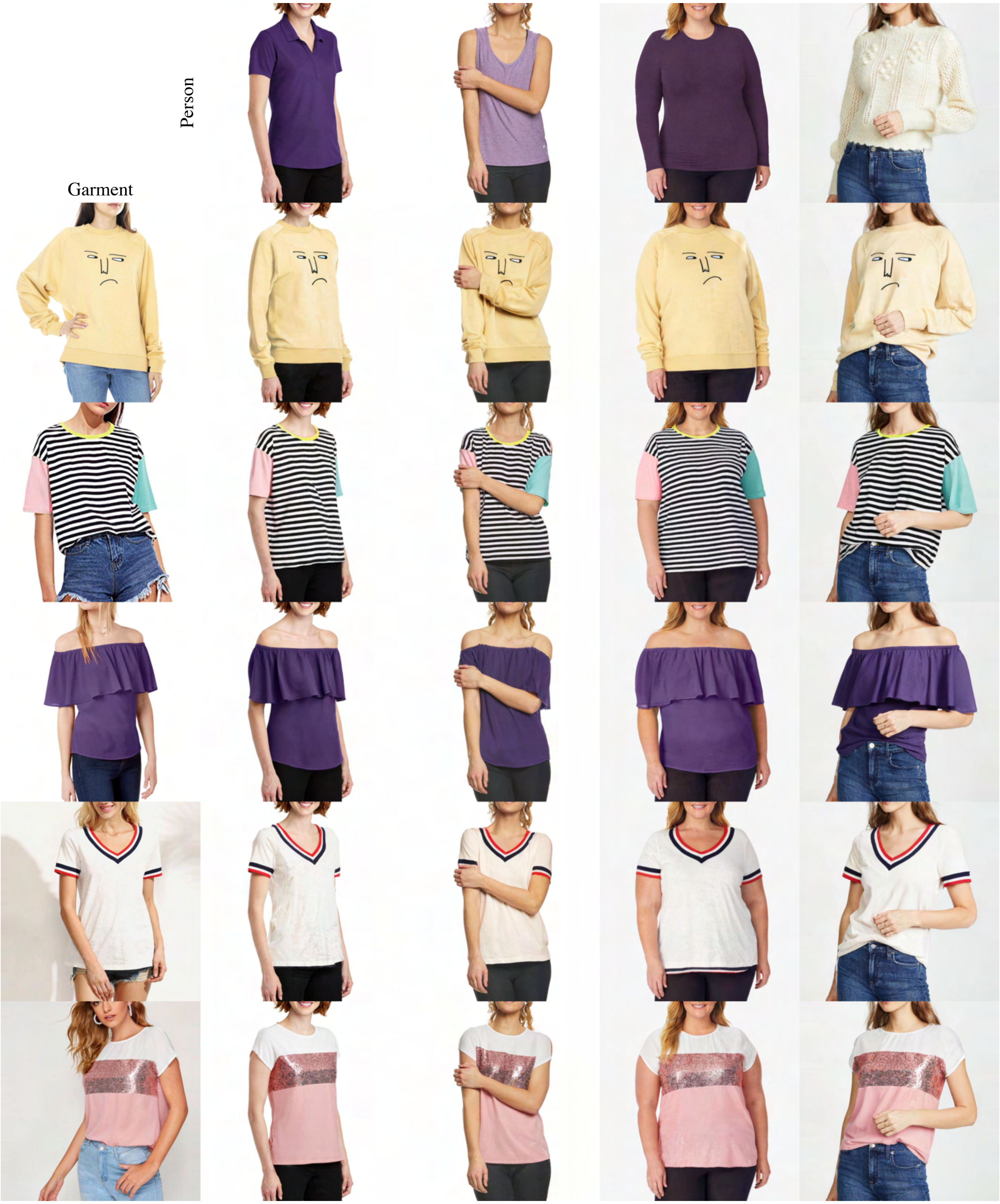

Figure 14. 4 women trying on 5 garments.

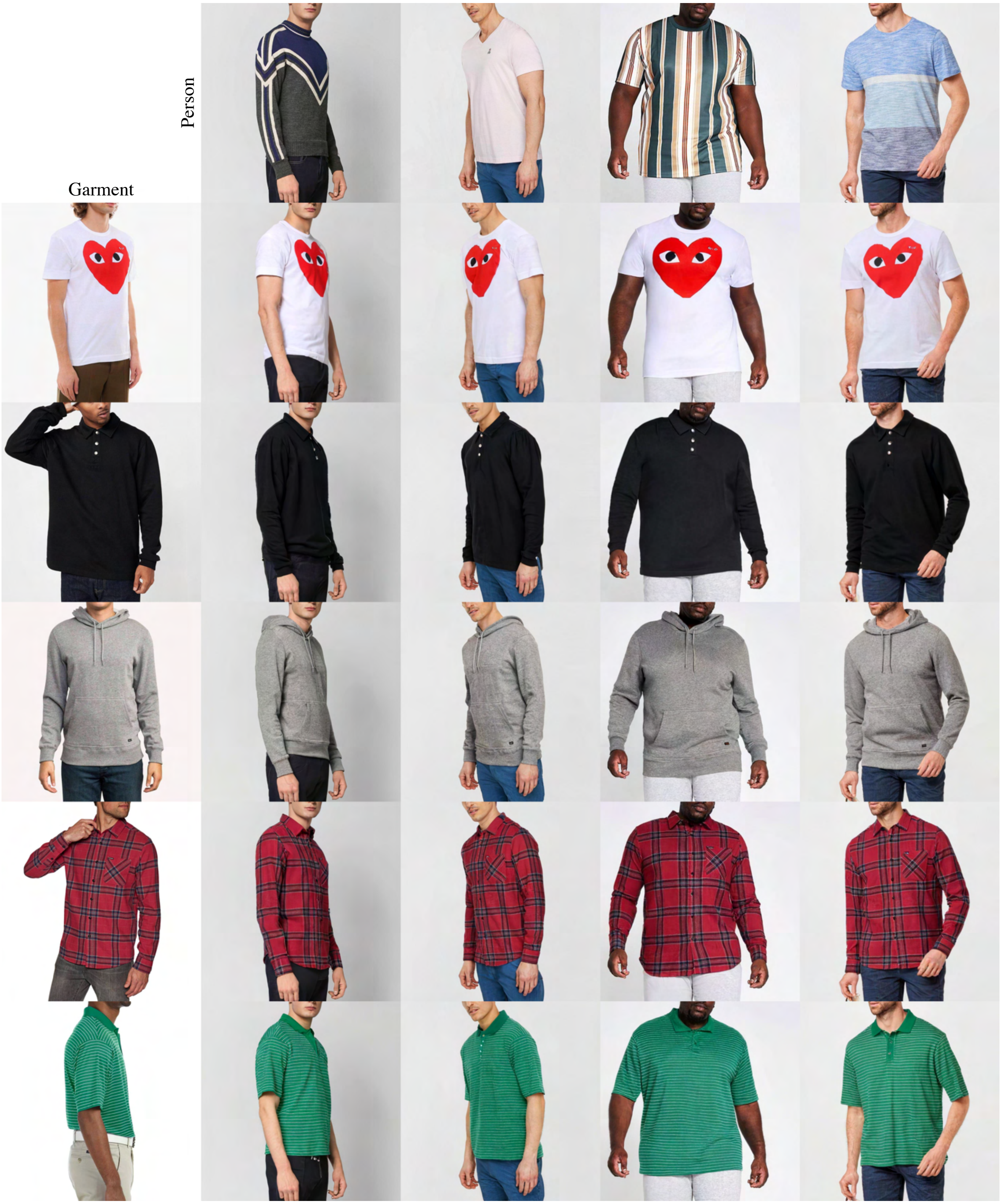

Figure 15. 4 men trying on 5 garments.

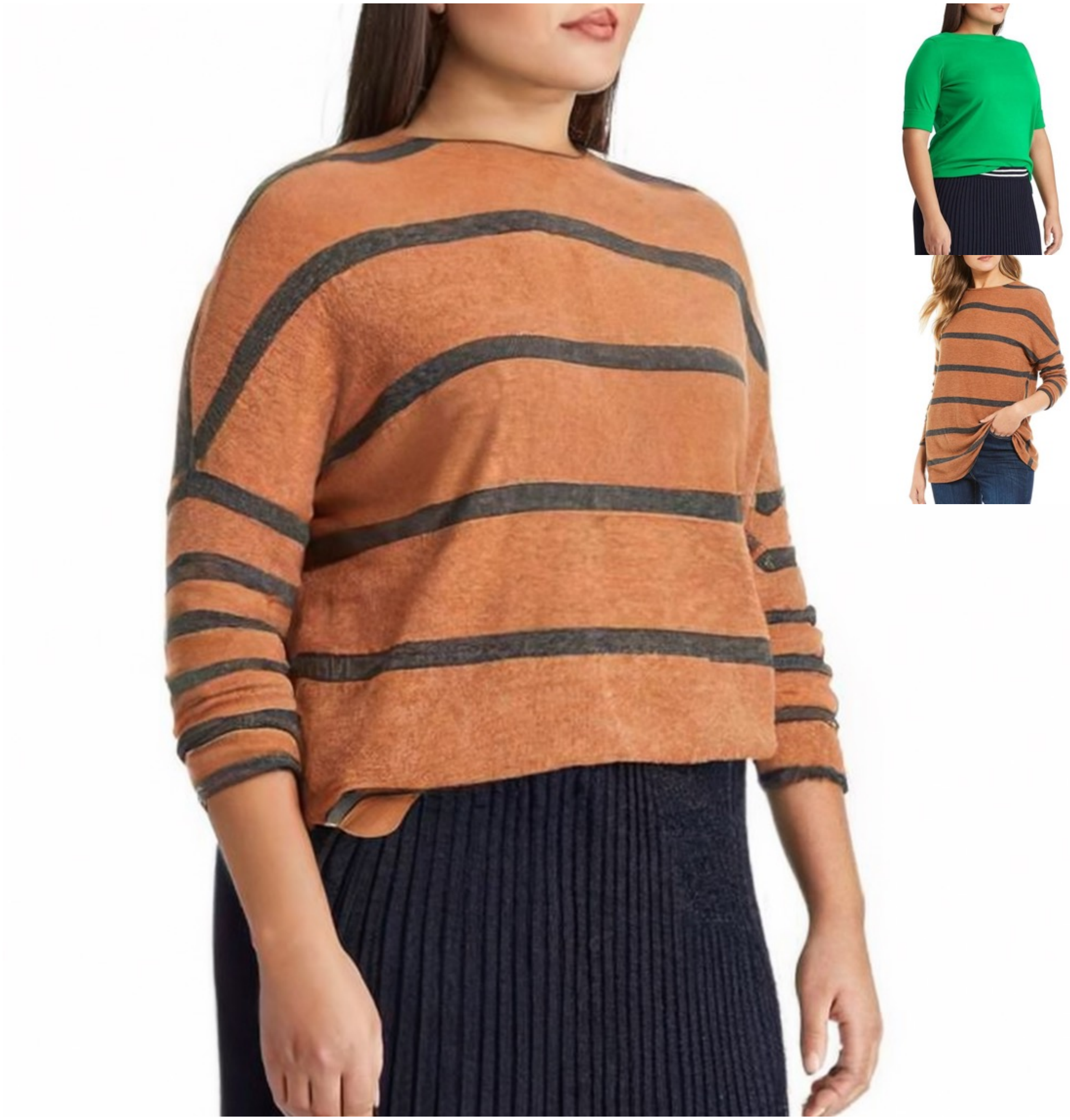

Figure 16. Larger version of teaser.

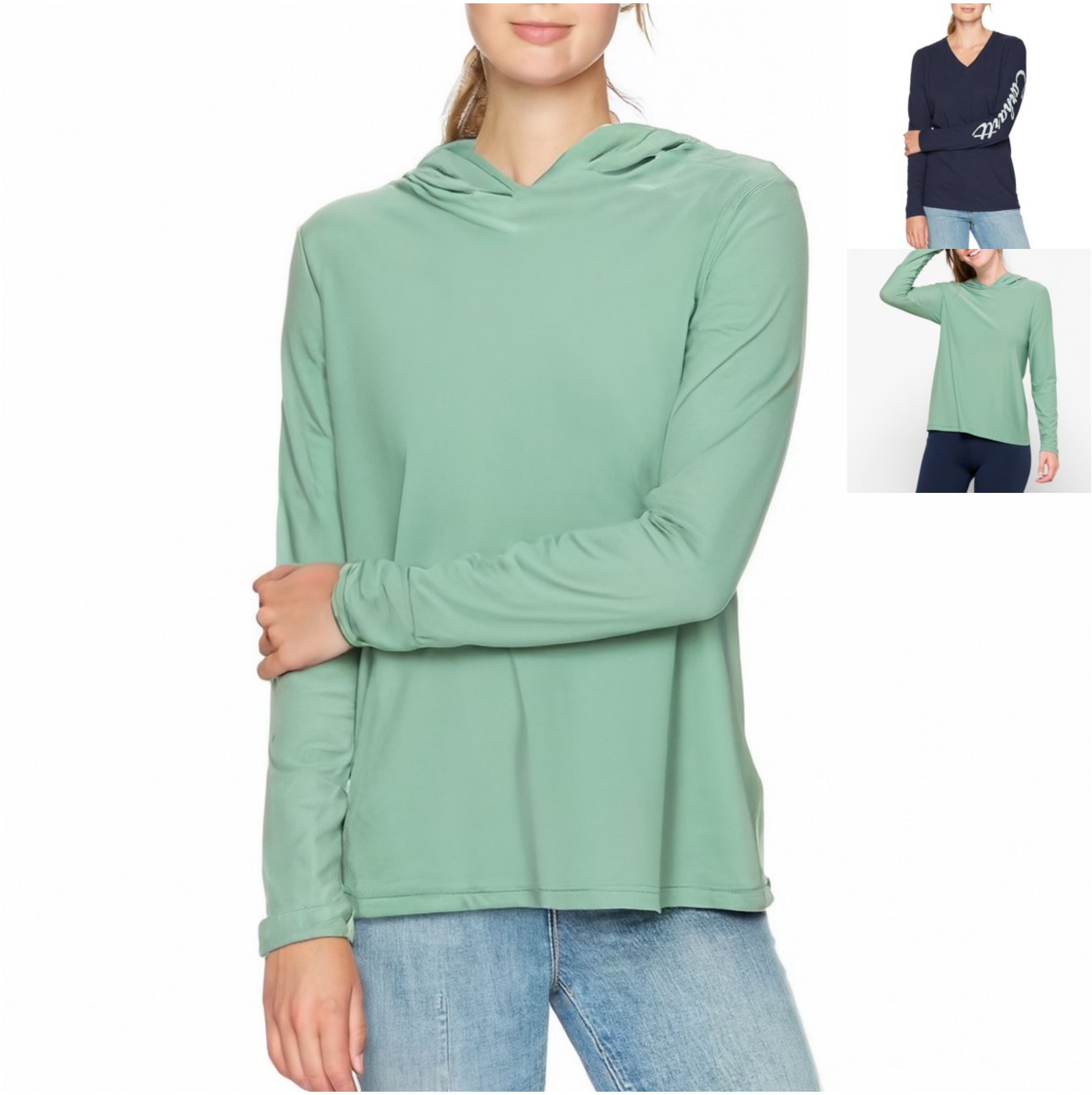

Figure 17. Larger version of teaser.

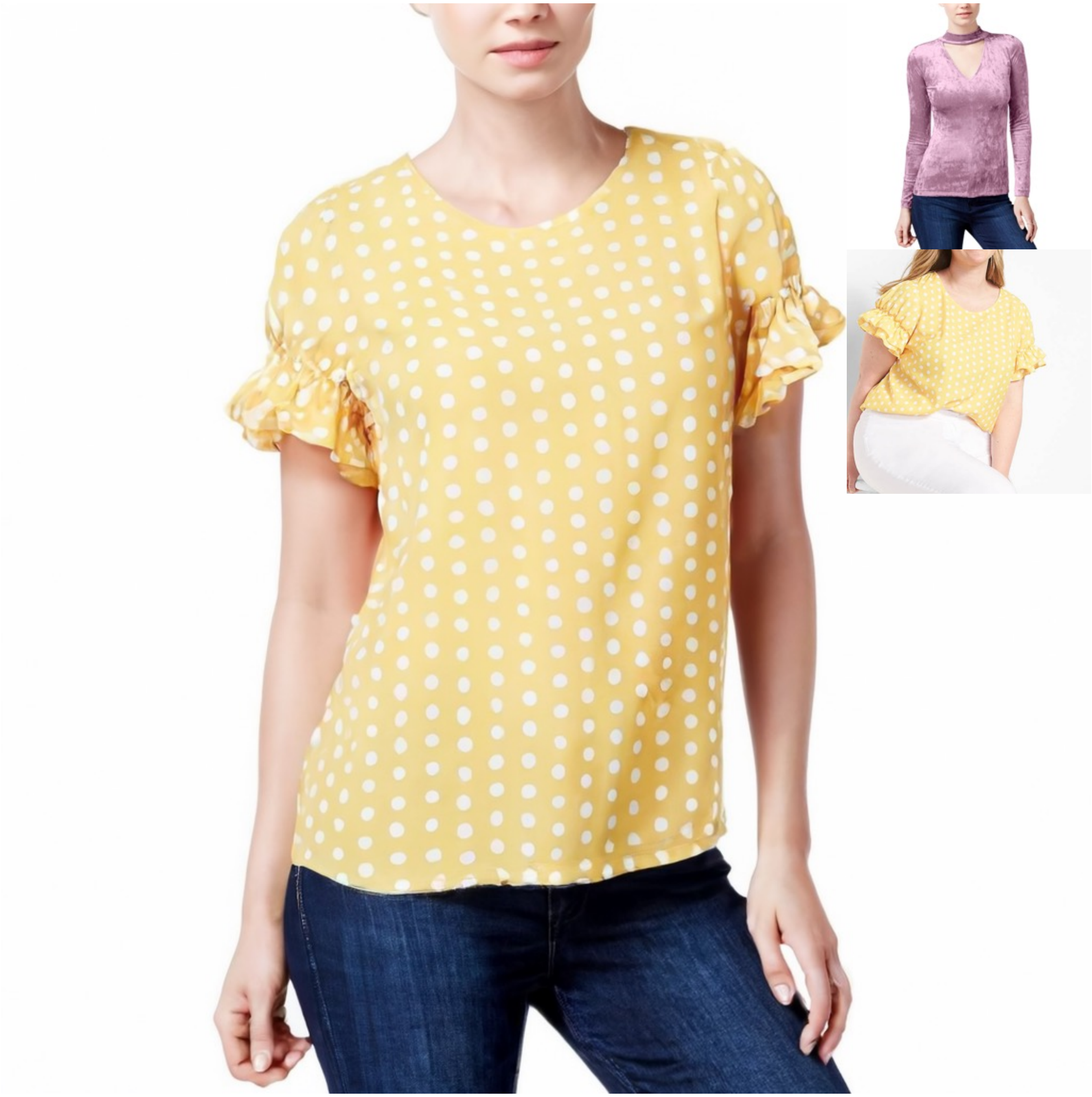

Figure 18. Larger version of teaser.

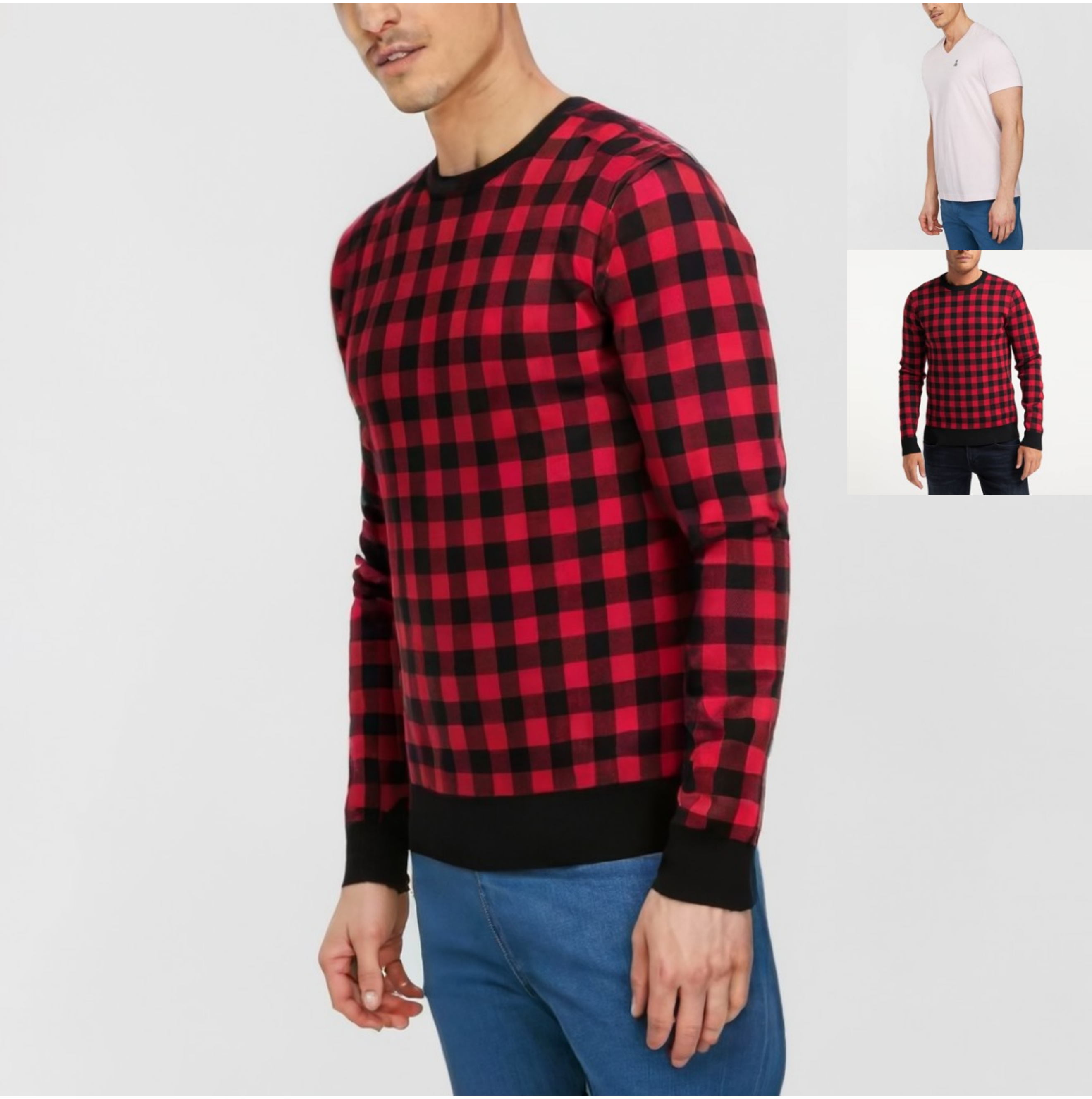

Figure 19. Larger version of teaser.

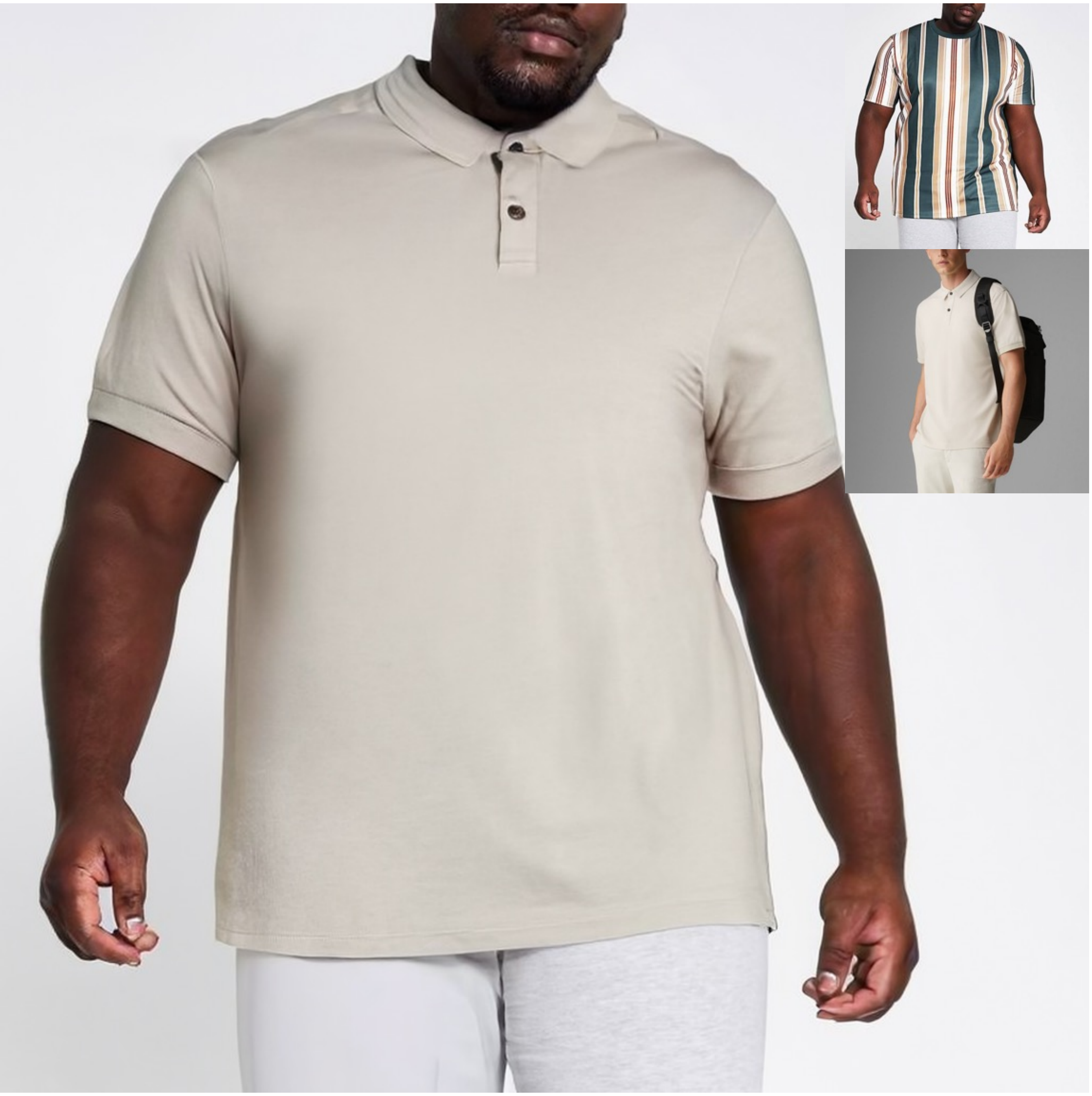

Figure 20. Larger version of teaser.

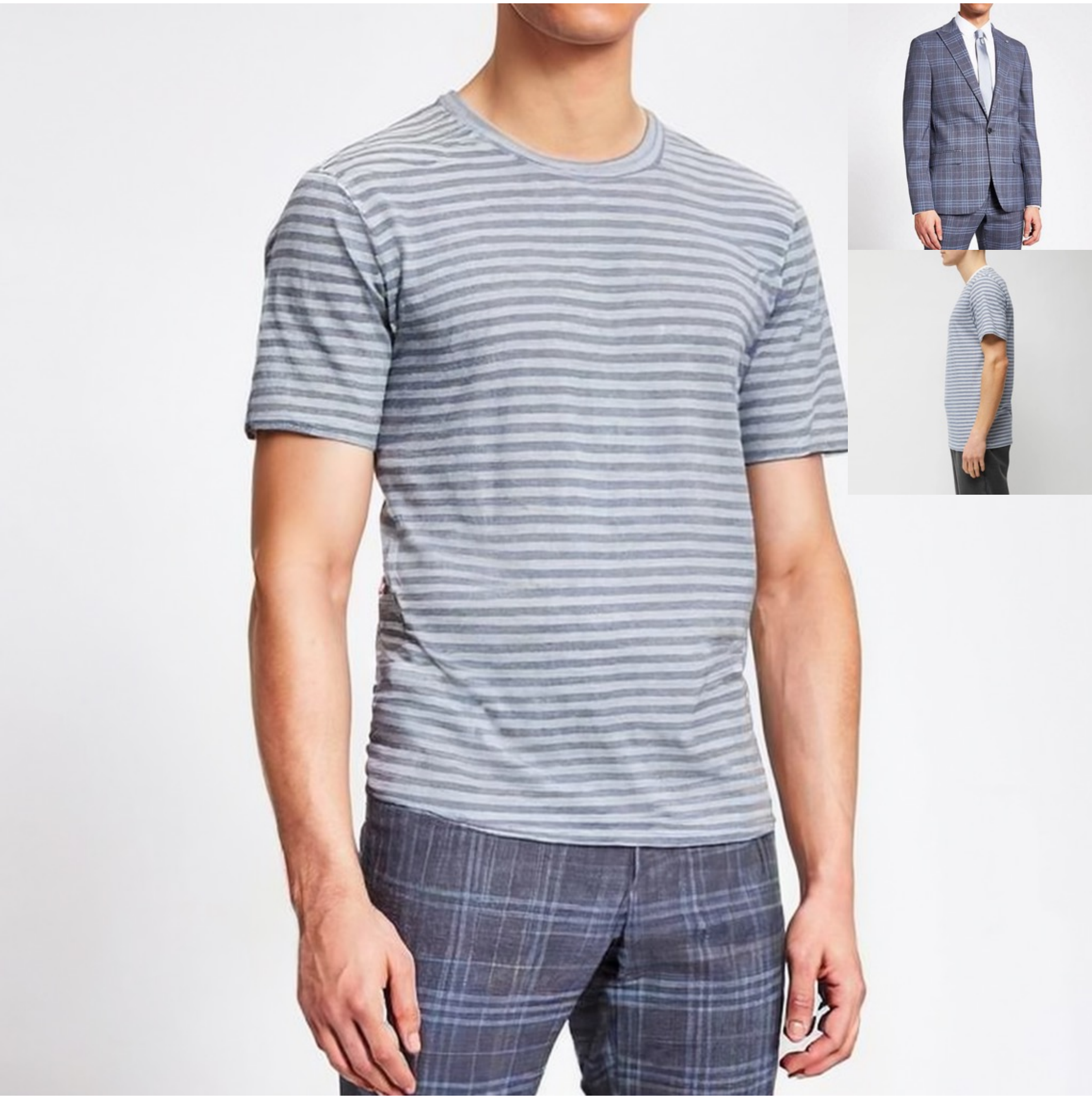

Figure 21. Larger version of teaser.