Artifact Does Matter! Low-artifact High-resolution Virtual Try-On via Diffusion-based Warp-and-Fuse Consistent Texture

Supplementary Material

1. Implementation Details

In both our *CTW* and *CTF* modules, the diffusion model settings are implemented with $T = 1000$ steps and a fixed variance schedule. We utilize the Adam optimizer with a learning rate of 1e-5. The batch size is set to 32. Additionally, in *CTW*, we set $\lambda_{clothes} = 0.2$, and $\lambda_{seg} = 1$. Images are resized to a 256×192 in *CTW*, and the output flow maps are up-resized then applied to the clothing images.

2. Additional Ablation Study

2.1. Ablation Study of CTW vs. CTF

We further qualitatively and quantitatively evaluate the effectiveness of our proposed *Conditional Texture Warping (CTW)* and *Conditional Texture Fusing (CTF)* by ablation study with GAN-based generators designed by HR-VITON [1]. Specifically, regarding the ablation study of *CTW*, we replace the *CTW* by the GAN-generated flow maps for warping clothes. For the ablation study of *CTF*, we replace the *CTF* by the GAN-based try-on generator to synthesize try-on results. The visual comparison Fig. 1 demonstrate that our *CTW* better stabilizes the warped texture, preventing clothing texture over-distortion (highlighted in red). Besides, our *CTF* fuses textures with consistency preventing unmatched color and texture degradation (highlighted in green). Moreover, Sec. 2.1 shows that our proposed LA-VTON with CTW and CTF surpasses the two ablation models replaced by GAN-based generators respectively in all 4 evaluation metrics, outperforming the ablation models by 64.5% in terms of Kernel Inception Distance (KID).

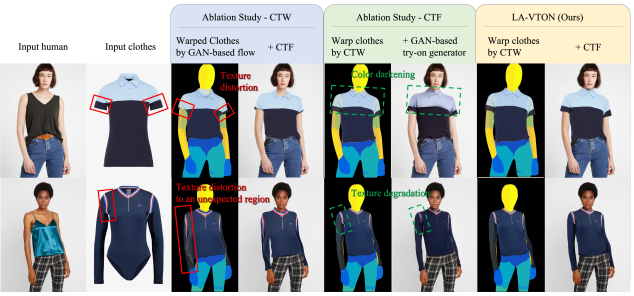

Figure 1. Ablation study of CTW and CTF.

Method	Paired		Unpaired	
	SSIM↑	LPIPS↓	FID↓	KID↓
X + *CTF*	0.851	0.136	12.06	0.329
CTW + Y	0.887	0.114	11.96	0.307
CTW + *CTF* (Ours)	**0.899**	**0.099**	**9.80**	**0.109**

NOTE: We describe the KID as a value multiplied by 100.

Table 1. Ablation study for *CTW* and *CTF*. X represents GAN-generated flow maps and Y is the GAN-based try-on generator.

2.2. Sampling Strategy

By using DDIM sampling, the reverse process can be performed in few steps. We analyze the effect of different sampling steps in *CTW* module by the warped clothes. In Fig. 2, we overlay the warped clothes onto the target person image to compare their alignment. The results manifest that using DDIM with $step = 1$ provides a rough alignment of the clothes with the person's shape. However, the patterns on the clothes appear distorted and cannot be preserved well. On the other hand, the results obtained with $step = 5$ and 10 preserve the clothing features well, and most areas are aligned accurately. Notably, the alignment is better for the cuffs in step size 10. To evaluate the performance of different steps, we calculated the IoU between the mask of warped clothes and the clothes region of the human image. The IoU for step sizes of 1, 5, 10, and 50 are 80.1%, 80.6%, 81.6%, and 82.1%, respectively. The IoU improves with an increase in DDIM steps, suggesting better alignment of details in the generated images. However, the improvements in IoU tend to plateau when the step size becomes larger, as the increase in steps may not result in significant quality improvements. Therefore, to strike a balance between image quality and computational efficiency, we set the number of steps to 10. This configuration allows the proposed method to produce satisfactory results while maintaining reasonable computational demands.

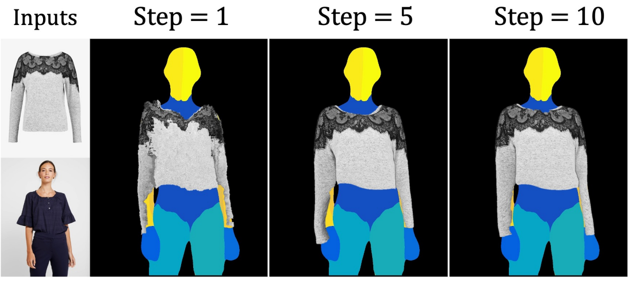

Figure 2. Comparison of different sampling steps.

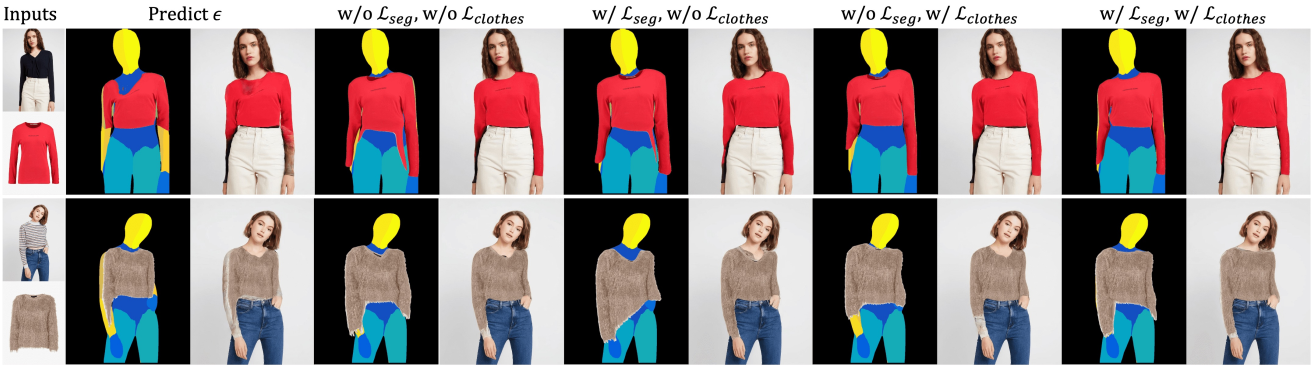
Inputs
Predict ϵ
w/o $\mathcal{L}_{seg}$, w/o $\mathcal{L}_{clothes}$
w/ $\mathcal{L}_{seg}$, w/o $\mathcal{L}_{clothes}$
w/o $\mathcal{L}_{seg}$, w/ $\mathcal{L}_{clothes}$
w/ $\mathcal{L}_{seg}$, w/ $\mathcal{L}_{clothes}$

Figure 3. **Ablation study on the effect of architecture design.** $\mathcal{L}_{clothes}$ helps align the warped clothes with the shape of the human body, while $\mathcal{L}_{seg}$ ensures the generated human segmentation corresponds to the warped clothes, resulting in good generation results. Both losses can improve the alignment of the warped results.

2.3. Training Objective

We conducted experiments to investigate the effectiveness of the training objective in *CTW* for clothing alignment in VITON-HD dataset. The results are summarized in Fig. 3 and Sec. 2.3. Firstly, we experimented with the objective of predicting noise ϵ instead of x_0. During training on higher-resolution images, the model predicting noise ϵ encountered stability issues and was prone to collapse, a phenomenon also reported in [2]. In contrast, training the model to predict x_0 maintained higher stability and better alignment in high resolutions. Therefore, we compared the model's performance of predicting noise ϵ at a lower resolution, which was 4× lower than our full model prediction. While it worked adequately, it led to obvious misalignment on arms, as shown in Fig. 3. Moreover, we experimented with different loss functions to assess their impact on clothing alignment, as illustrated in Fig. 3. In addition, Sec. 2.3 presents the quantitative results of using different training losses. Specifically, training with only $\mathcal{L}_{flow}$ resulted in rough alignment of the clothes, while adding $\mathcal{L}_{clothes}$ and $\mathcal{L}_{seg}$ significantly improved all metric scores. Using both losses together achieved the best performance in terms of clothing alignment and overall image quality.

Method	Paired		Unpaired	
	SSIM↑	LPIPS↓	FID↓	KID↓
Ours (predict ϵ)	0.892	0.137	11.91	0.332
Ours (w/o $\mathcal{L}_{seg}$, $\mathcal{L}_{clothes}$)	0.893	0.134	11.31	0.272
Ours (w/o $\mathcal{L}_{clothes}$)	0.892	0.137	11.42	0.281
Ours (w/o $\mathcal{L}_{seg}$)	0.892	0.135	11.40	0.290
Ours	**0.899**	**0.099**	**9.79**	**0.109**

NOTE: We describe the KID as a value multiplied by 100.

Table 2. Quantitative comparison for different training objectives.

2.4. Sensitivity Analysis for Loss Weights

For the hyper-parameters $\lambda_{clothes}$ and λ_{seg} corresponding to the designed losses in our *CTW* module, we conducted experiments involving a multiplication factor of 100 to assess their sensitivity. The quantitative results are illustrated in Sec. 2.4, while the qualitative outcomes are presented in Fig. 4. The results reveal that when $\lambda_{clothes}$ is excessively large, distortion in clothing occurs. This distortion arises due to pixel-wise loss causing significant gradients in the model, making it challenging for the model to learn the original distribution of flow, and failing to preserve the original texture. On the other hand, if λ_{seg} is too large, the model focuses more on generating the segmentation map and fails to learn the accurate warping, and further misses the alignment between warped clothes and the segmentation map. Hence, we set $\lambda_{seg} = 1$ and $\lambda_{clothes} = 0.2$ for our full model to have the best quality.

$\lambda_{clothes}$	λ_{seg}	Paired		Unpaired	
		SSIM↑	LPIPS↓	FID↓	KID↓
20	1	0.838	0.156	11.89	0.258
0.2	100	0.843	0.139	10.43	0.160
0.2	1	**0.899**	**0.099**	**9.79**	**0.109**

NOTE: We describe the KID as a value multiplied by 100.

Table 3. Quantitative comparison for different loss weights.

2.5. Training of Conditional Texture Fusing module

In *Conditional Texture Fusing module*, we employed a scheme where the clothing image is multiplied by the clothes masks to help the model address the issue of misalignment in warped clothes. In the full model, the scheme

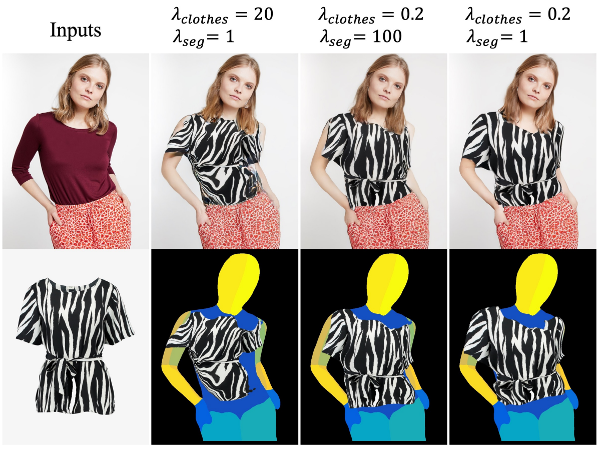

Figure 4. Comparison of different λ settings in CTW module.

of C_{cond} is derived as follows:

$$Train : C_{cond} = I_c \odot \mathcal{W}(C_{mask}, \hat{x_0}), \tag{1}$$

$$Test : C_{cond} = C_{warp} \odot \hat{S}_c, \tag{2}$$

To demonstrate the effectiveness of this masking scheme, we conducted experiments without this scheme, *i.e.*,

$$Train : C_{cond} = I_c, \tag{3}$$

$$Test : C_{cond} = C_{warp}. \tag{4}$$

As shown in Fig. 5, when the model is trained and tested without the mask, the generated clothing appears artifacts along the edges (the red circle) whenever the clothes are slightly misaligned. Additionally, the body parts of the generated results are also unnaturally occluded (the blue circle) when the clothes go beyond the intended region.

3. Occlusion Results

Fig. 6 illustrates cases of occlusion, demonstrating that our model is capable of generating good results even when the arms obstruct the clothing.

4. Failure Cases

Failure cases of our model are usually caused by complex poses in target human images or incorrect clothing masks. To illustrate the failure examples, we provide the following:
Complex Pose. As Fig. 7 shows, artifacts occur when the target person is in complex poses. When there are large movements in the input person, such as raising their hands above their heads, our *CTW* module tends to produce incorrect warping, leading to artifacts in the output image. The reason is that complex poses are very rare in the VITON-HD dataset, making it difficult for the model to learn effectively with limited data. We will tackle this issue in future developments.

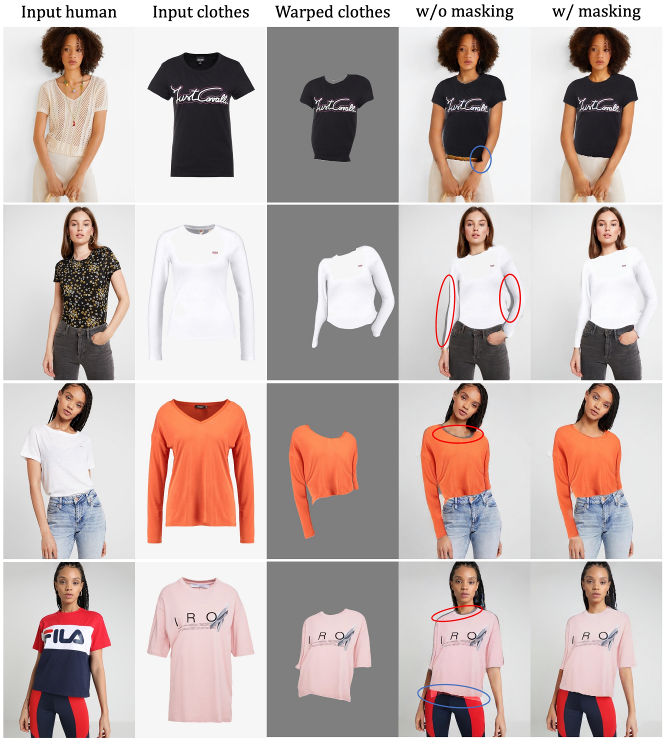

Figure 5. **Comparison of different C_{cond} in CTF module.** The column of w/o masking is the results of using only I_c and C_{warp} as C_{cond} in training and inference time.

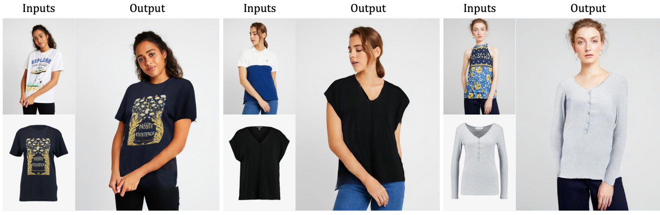

Figure 6. Occlusion cases.

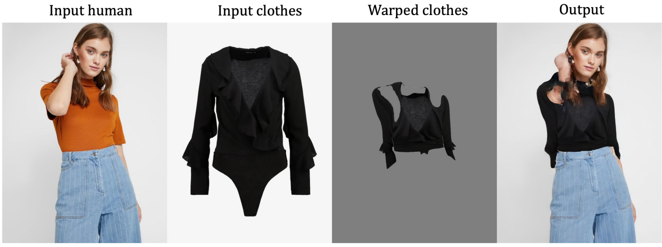

Figure 7. Failure cases of complex poses.

Incorrect Mask. As Fig. 8 shows, when predicted clothing masks are failed, it leads to failure try-on results. The clothing mask in the data may sometimes be inaccurate, especially when the color of the clothes is too similar to the background. Incorrect clothing masks make it hard for the model to accurately recognize the shape of the clothes, leading to erroneous warping and incorrect generation results. HR-VITON mentions the use of a discriminator to handle such cases, but this problem has not been fundamentally resolved. We also look forward to optimizing and resolving this issue in future method designs.

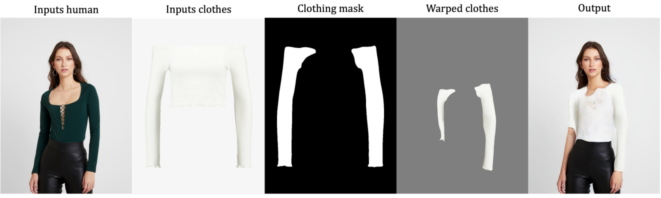

Figure 8. Failure cases of incorrect masks.

5. Additional Qualitative results

We provide additional qualitative comparisons in Figs. 9 to 14. Fig. 9 shows that our method outperforms others in generating low artifact results even in simple clothing types, *e.g*., plain color thin strap vests, T-shirts, and shirts. Meanwhile, Figs. 10 to 12 demonstrate our method's efficacy in preserving clothing shape with complex decorations, *e.g*., puff sleeves, cross-strap vests, turtleneck shirts, etc. Specifically, the example in Fig. 10 highlights our ability to preserve the special shape of sleeves, *e.g*., puff sleeves, shoulder pad on T-shirt, and text/line design on side arm. Fig. 11 includes side bow tie design and cross-strap vests, which are rare clothing styles in the VITON-HD dataset. Our method accurately generates try-on results for these special designs. Fig. 12 demonstrates the performance on preserving patterns around the neckline and bottom of clothes. Additionally, Figs. 13 and 14 showcase the outperforming texture-preserving capabilities of our method.

References

[1] 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

[2] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

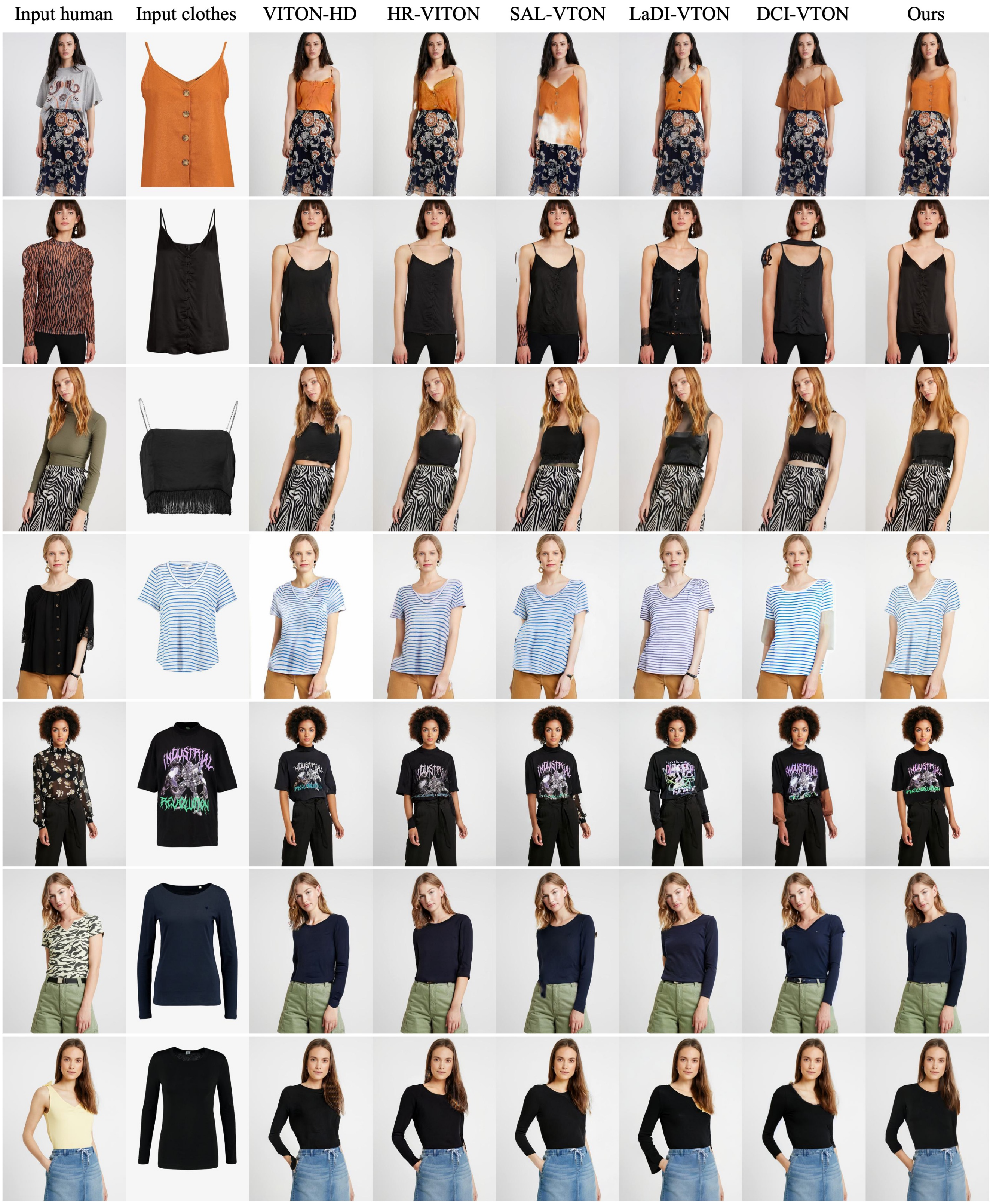

Figure 9. Additional comparison with state-of-the-art try-on methods. Our method outperforms others in generating low artifact results even in simple clothing types, *e.g*., plain color thin strap vests, T-shirts, and shirts.

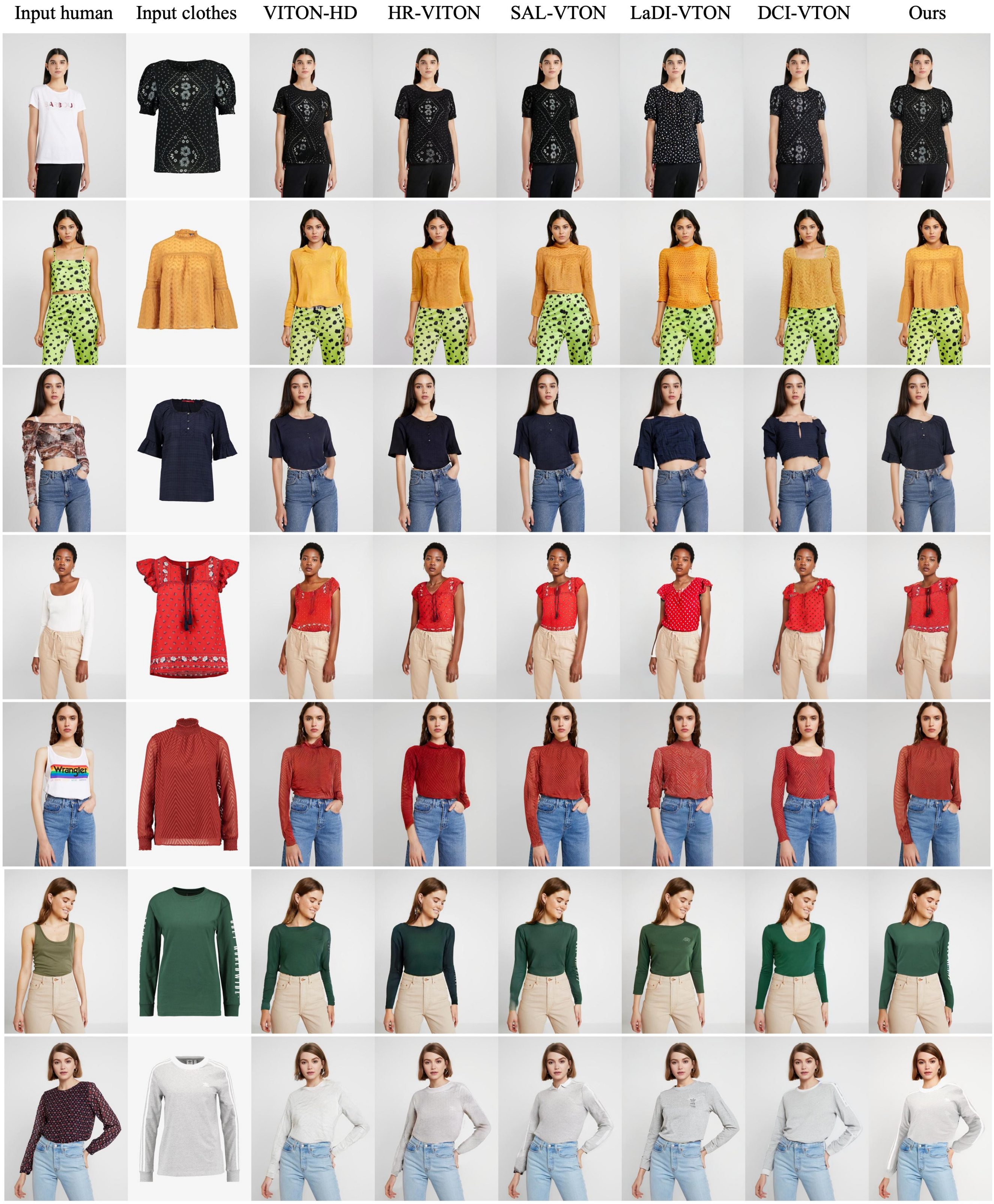

Figure 10. Additional comparison with state-of-the-art try-on methods. It highlights our ability to preserve the special shape of sleeves, *e.g.*, puff sleeves, shoulder pad on T-shirt, and text/line design on side arm.

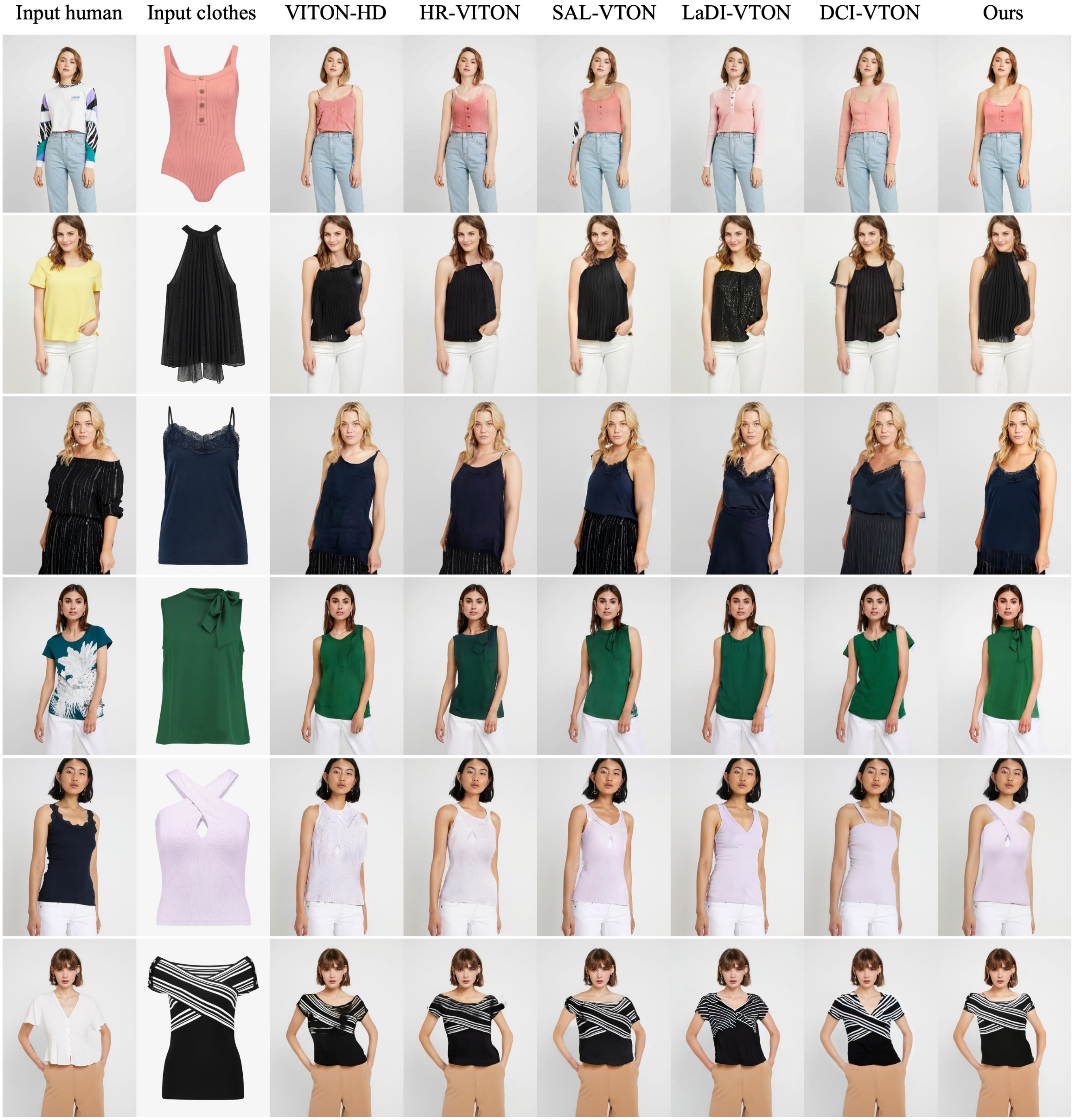

Figure 11. Additional comparison with state-of-the-art try-on methods. Our method accurately generates try-on results for side bow tie designs, cross-strap vests, which are rare clothing styles in the VITON-HD dataset.

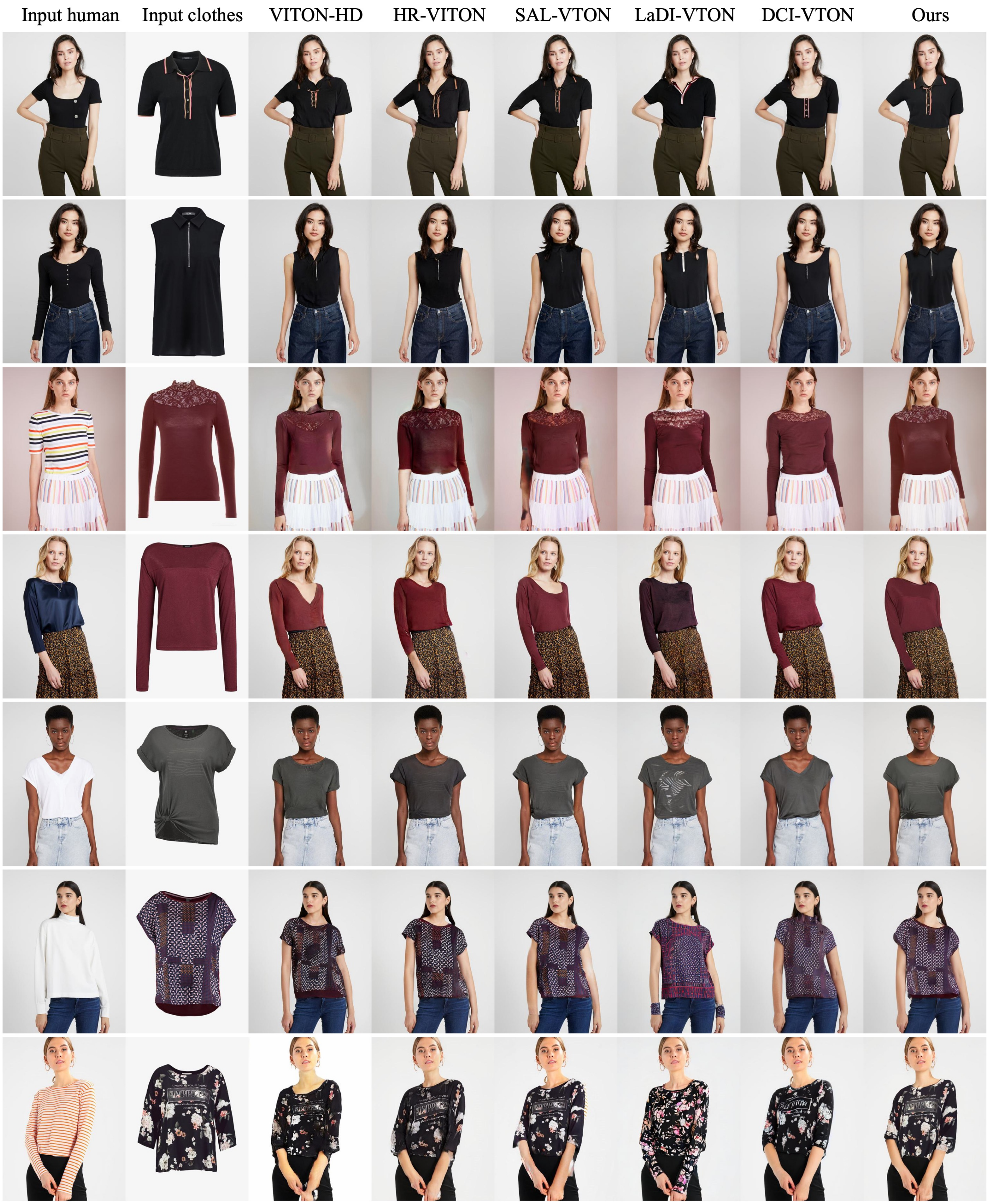

Figure 12. Additional comparison with state-of-the-art try-on methods. It demonstrates the performance on preserving patterns around the neckline and bottom of clothes.

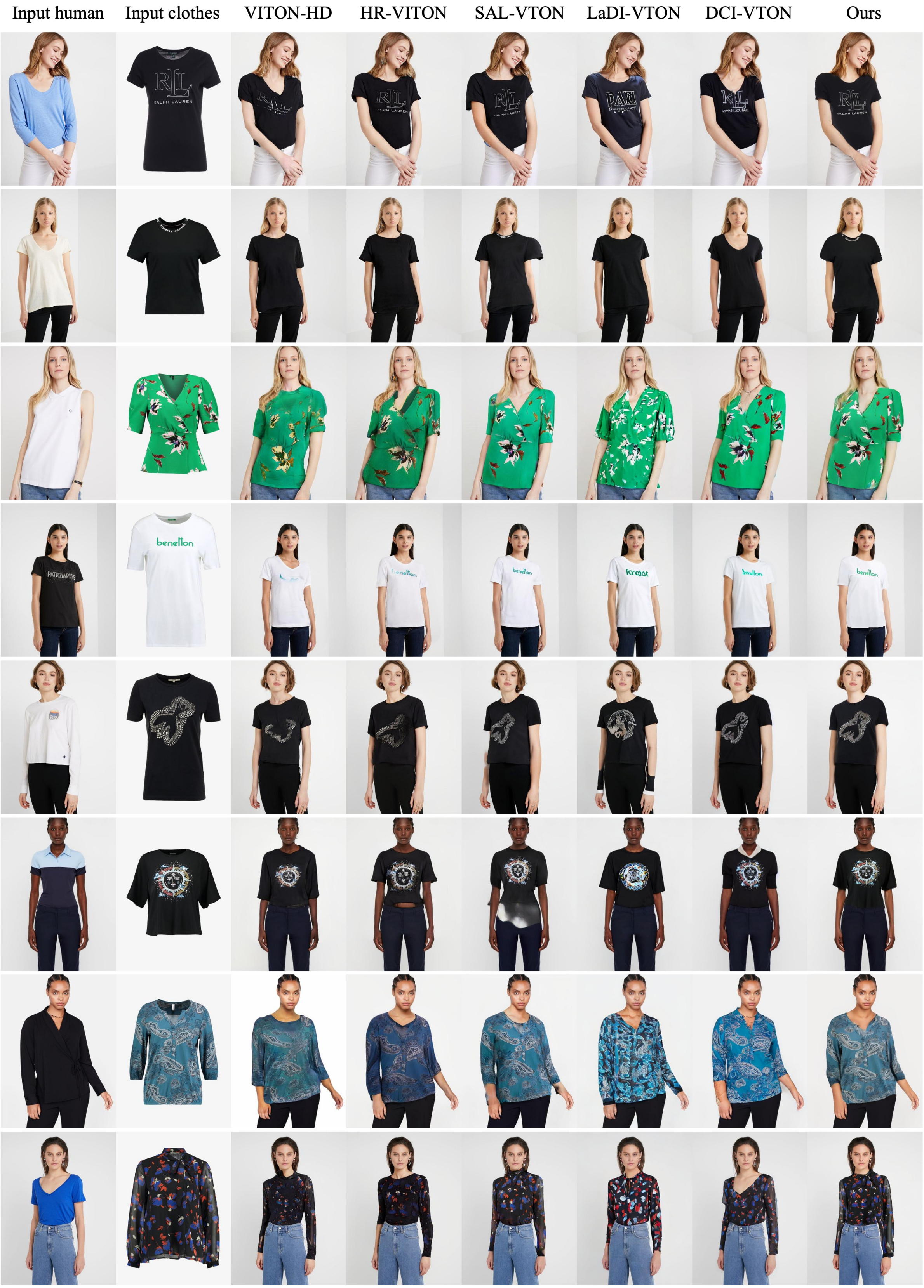

Figure 13. Additional comparison with state-of-the-art try-on methods. Our method showcase the outperforming texture-preserving capabilities.

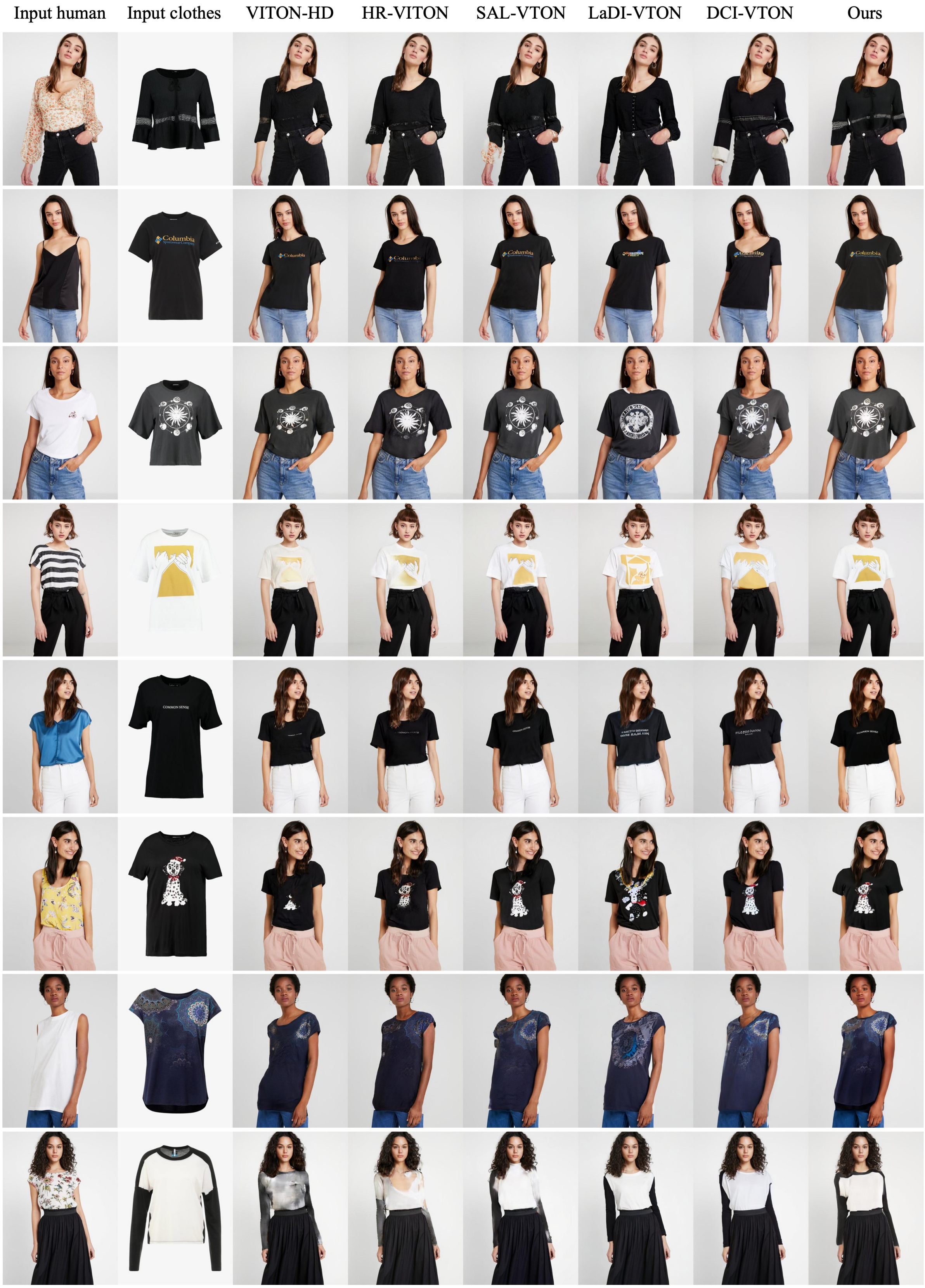

Figure 14. Additional comparison with state-of-the-art try-on methods. Our method showcase the outperforming texture-preserving capabilities.