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

Chiang Tseng Chieh-Yun Chen Hong-Han Shuai
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
{chiang.ee11, astrid, hhshuai}@nycu.edu.tw

Figure 1. Our method outperforms the existing SOTAs in generating low-artifacts images while preserving garment details.

Abstract

In virtual try-on technology, achieving realistic fitting of clothing on human subjects without sacrificing detail is a significant challenge. Traditional approaches, especially those using Generative Adversarial Networks (GANs), often produce noticeable artifacts, while diffusion-based methods struggle with maintaining consistent texture and suffer from high computational demands. To overcome these limitations, we propose the Low-artifact High-resolution Virtual Try-on via Diffusion-based Warp-and-Fuse Consistent Texture (LA-VTON). This novel framework introduces Conditional Texture Warping (CTW) and Conditional Texture Fusing (CTF) modules. CTW improves warping stability through simplified denoising steps, and CTF ensures texture consistency and enhances computational efficiency, achieving inference times $17\times$ *faster than existing diffusion-based methods. Experiments show that LA-VTON surpasses current SOTA high-resolution virtual try-on methods in both visual quality and efficiency, marking a significant advancement in high-resolution virtual try-on and setting a new standard in digital fashion realism.*

1. Introduction

Virtual try-on technology aims to seamlessly integrate clothing onto a target individual's image. This task has been propelled by rapid advancements in generative AI, leading to a surge of research across various methodologies, from image-based, single-pose methods to dynamic, multi-pose, and video-based systems [1–7, 9, 11–13, 15, 16, 20–24].

There are two main streams in virtual try-on methods: GAN-based and Diffusion-based virtual try-on. Regarding GAN-based virtual try-on, VITON-HD [3] pioneered high-resolution virtual try-on by introducing misalignment-aware normalization to address texture discrepancies. Moreover, HR-VITON [13] designed a dual-path method to simultaneously synthesize warped clothes and human segmentation, leading to structurally aligned clothes. Yet, both VITON-HD and HR-VITON exhibit notable shortcomings in properly warping clothing to fit the target body shapes, as highlighted in Fig. 1. This issue arises from the substantial spatial discrepancy between the original clothing items and their intended placement on the human figure, making it difficult to warp clothes effectively in one step. On the diffusion front, models like LDM [18] and SDXL [17] have

set benchmarks in image synthesis tasks, but often struggle to generate consistent and detailed patterns in virtual try-on, particularly when conditioned on image inputs rather than textual prompts. Recent attempts to leverage diffusion models for virtual try-on [9, 16, 24] have made progress but continue to wrestle with rendering precise text and intricate textures on the clothing, as illustrated in Figs. 1 and 5, especially synthesizing non-existent textures. Moreover, these diffusion-based generators often suffer from lengthy inference times, especially for high-resolution outputs.

To address these challenges, we propose a Low-artifact High-resolution Virtual Try-on via Diffusion-based Warp-and-Fuse Consistent Texture (LA-VTON). Specifically, to surmount the challenges of garment warping and texture consistency in high-resolution virtual try-on, we first design the diffusion-based *Conditional Texture Warping Module (CTW)*, offering a novel alternative that breaks down the complex warping task into a sequence of simpler, more controlled denoising steps. This novel approach aims to enhance texture stabilization, mitigating the risk of over-distortion and ensuring more consistent patterns. Yet, despite achieving more accurately aligned warped clothing, the task of flawlessly integrating these textures within the final synthesized image continues to present challenges. As indicated in Fig. 1, even with structurally aligned garments, all baselines encounter texture fidelity difficulties. VITON-HD and HR-VITON exhibit pattern degradation and darkening. These phenomena indicate *artifact collapse*.[1] LaDI-VTON and DCI-VTON struggle to generate consistent clothing textures. To solve these issues, we introduce the *Conditional Texture Fusing Module (CTF)*. This module reconfigures the latent diffusion model to exclude cross-attention, enabling direct high-resolution try-on synthesis of high quality with 17× inference time acceleration.

Our contributions are summarized as follows. First, we develop a 2-stage diffusion-based virtual try-on method in high-resolution (1024×768), which addresses GAN-based try-on issues, i.e., warping misalignment and texture inconsistency, and diffusion-based try-on issues, i.e., texture inconsistency and time-consuming. Second, we propose the novel *Conditional Texture Warping Module* to ensure clothing warping stability, preventing misalignment and texture over-distortion. Subsequently, we design the effective *Conditional Texture Fusing Module* to seamlessly fuse human and clothing textures. Finally, extensive experiments show that our model significantly outperforms SOTAs, with at least 36.25% improvement in terms of KID.

[1]This occurs when the model repeatedly introduces the same types of errors or distortions across different outputs. These artifacts might manifest as specific patterns, textures, or anomalies that are not present in the training data and are consistently reproduced in the generated results, e.g., color darkening shown in Fig. 1 and Fig. 5.

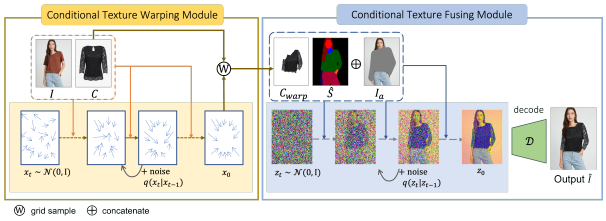

Figure 2. Overview of our framework.

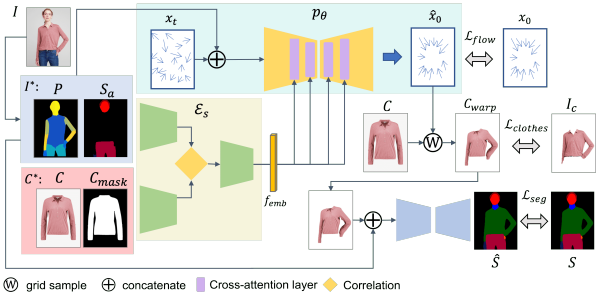

Figure 3. Architecture of *Conditional Texture Warping Module*.

2. Proposed Method: LA-VTON

Fig. 2 shows the the architecture of our proposed LA-VTON, comprising two main components: the *Conditional Texture Warping Module* and the *Conditional Texture Fusing Module*, both rely on diffusion models as their core. In the first module, an appearance flow map is generated using an implicit diffusion model. This flow then enables the transformation of the clothing image C into a warped clothing image C_{warp} aligned with human image I. Subsequently, the second module integrates the human information with C_{warp} to generate the try-on result. In the following, we discuss the details of the LA-VTON framework.

2.1. Conditional Texture Warping Module

To address the perceptual issues present in the warping methods of prior virtual try-on tasks [3, 13], we introduce an innovative diffusion-based *Conditional Texture Warping Module (CTW)*, which more effectively aligns clothes with the target body shape while preserving texture. Fig. 3 illustrates the architecture of *CTW*, which focuses on two main objectives: (i) warping clothes with consistent texture, and (ii) predicting human segmentation to enhance warped clothes alignment.

Diffusion-based Clothing Deformation. In this stage, we train a conditional diffusion model $p_\theta(x|I, C)$, where the result x, representing the appearance flows, should accurately warp clothing image C to fit human image I while maintaining the inner texture consistency. To represent the structure of human image I, we propose to utilize both the human dense pose P (derived by [10]) and clothing-agnostic human segmentation S_a (derived by [8]). The encoder $\mathcal{E}_s$ first extracts features f_{emb} from the human struc-

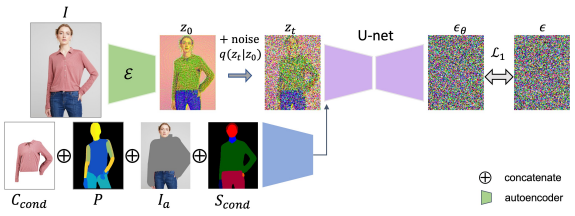

Figure 4. Architecture of *Conditional Texture Fusing Module*.

ture $I^* = [P, S_a]$ and the clothing $C^* = [C, C_{mask}]$, where C_{mask} represents the mask of the in-shop clothes C.

To transmit rich structural information from the source image to the model, we transfer f_{emb} through cross-attention, and also concatenate I^* with the denoised input for better alignment. This allows the network to fully exploit the correspondences between the human and clothing structure, thus resulting in low-distortion appearance flows.

As shown in Fig. 2, the diffusion process follows the objective proposed by [19]. To improve stability during training [14], we train our model to predict x_0 instead of noise. We adopt flow loss $\mathcal{L}_{flow}$ to learn x_0 from paired I and C:

$$\mathcal{L}_{flow} = \|p_\theta(x_t, t, I^*, C^*) - x_0\|. \quad (1)$$

Besides, to better align the clothing texture, clothes loss $\mathcal{L}_{clothes}$ is applied to the warped clothes:

$$\mathcal{L}_{clothes} = \|C_{warp} - I_c\|, \quad C_{warp} = \mathcal{W}(C, \hat{x_0}) \quad (2)$$

where $\mathcal{W}$ represents the grid sampling from source image C in terms of predicted flow $\hat{x_0}$, and I_c represents the clothing region of the human image I.

Human Segmentation. We employ an additional U-Net to predict human segmentation for two purposes: (i) enhancing the alignment between the warped clothes and the human body, and (ii) providing guidance for synthesizing try-on results in the next stage. The model concatenates I^* and C_{warp} as inputs to predict human segmentation $\hat{S}$, where we utilize focal loss $\mathcal{L}_{seg}$ as supervision.

Overall, the *Conditional Texture Warping Module*, including a diffusion model, a texture encoder, and a segmentation prediction U-Net, is trained in an end-to-end manner. Therefore, the loss of predicted human segmentation is propagated back to the diffusion model and makes the warped clothes and the clothing channel of $\hat{S}$ synchronous structurally. The overall loss function of *CTW* is:

$$\mathcal{L}_{CTW} = \mathcal{L}_{flow} + \lambda_{clothes}\mathcal{L}_{clothes} + \lambda_{seg}\mathcal{L}_{seg}, \quad (3)$$

where $\lambda_{clothes}$ and λ_{seg} are hyperparameters.

2.2. Conditional Texture Fusing Module

After addressing the challenge of clothing alignment, we further improve the visual quality of final try-on results.

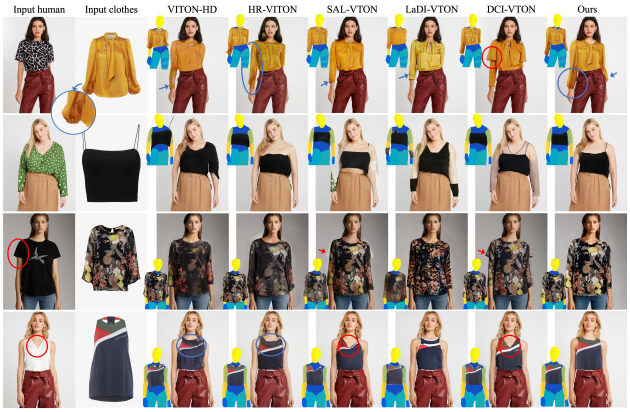

Figure 5. Visual comparison with SOTA baselines.

As shown in Fig. 1, other methods fail to preserve the texture even in simple human posture. Accordingly, we propose the *Conditional Texture Fusing Module* to leverage the strengths of diffusion models on synthesizing realistic images while addressing its limitations on texture consistency.

As depicted in Fig. 4, the human image $I \in \mathbb{R}^{3\times H\times W}$ is first encoded to the latent space $z = \mathcal{E}(I) \in \mathbb{R}^{4\times h\times w}$ using a pre-trained autoencoder from [18] to accelerate the learning process since the computational complexity of the sampling from the true posterior distribution is reduced. Besides, since the diffusion process is performed under latent space, we employ a condition encoder to extract latent features from I_{cond}, which consists of clothing condition C_{cond} and human segmentation S_{cond} derived from *CTW*, clothing-agnostic human image I_a and dense pose P, and then concatenate to U-net to fuse target clothing features to human. Thanks to the aligned clothing conditions from the previous stage, we exclude the cross-attention layer in U-net and takes the conditions only through concatenation. This approach significantly reduces computational complexity. At this stage, the model is learned to predict noise ϵ.

$$\mathcal{L}_{simple} = \mathbb{E}_{x_0,t,\epsilon}[\|\epsilon - \epsilon_\theta(z_t, t, I_{cond})\|_2^2]. \quad (4)$$

3. Experiments

3.1. Experimental Setup

Dataset. We train and evaluate our proposed LA-VTON on the VITON-HD dataset [3], which comprises 13,679 high-resolution (1024×768) frontal-view images of women wearing tops, along with corresponding clothing items. The dataset are split into 11,647/2,032 for training/testing pairs.

Baselines. We compare our proposed LA-VTON with SOTAs on VITON-HD dataset, which includes three GAN-based methods: VITON-HD [3], HR-VITON [13] and SAL-VTON [21], and two LDM-based approaches: LaDI-VTON [16] and DCI-VTON [9]. We use the official codes provided by the respective authors to obtain baseline results.

3.2. Qualitative Results

As presented in Figs. 1 and 5, LA-VTON achieves visually convincing high-resolution try-on results, ensuring both clothing warping stability and texture fusing consistency.

VITON-HD employs a TPS-based warping method, resulting in significantly misaligned warped clothes, particularly when trying on clothes with complex logos as shown in the second row in Fig. 1. Meanwhile, HR-VITON devises a dual-path method to simultaneously synthesize warped clothes and human segmentation, leading to structurally aligned clothes. On the other hand, SAL-VTON incorporates additional landmarks for precise warping. However, as depicted in Fig. 5, they struggle to effectively maintain the shape of the clothing. For instance, none of the baselines can preserve flared sleeves in the first row or tube tops in the second row, as indicated by blue marks. In contrast, our designed *CTW* enhances the warping process, ensuring highly preserved clothes shape. Additionally, VITON-HD and HR-VITON would destroy the clear color texture due to *artifact collapse* as the third and fourth rows show.

Regarding LaDI-VTON and DCI-VTON, they often produce try-on results with inconsistent styles and textures, because they rely on the large pre-trained diffusion models as their backbone for synthesis. The cross-attention mechanism for global conditions input can lead to inconsistency in garment details. In contrast, our proposed *CTF* can harness the generative capabilities of the diffusion model while ensuring the preservation of details in the warped clothes.

3.3. Quantitative Results

Our evaluation includes both objective metrics and a subjective user study. The evaluation metrics include *SSIM*, *LPIPS*, *FID*, and *KID*, which are commonly used in virtual try-on tasks. The user study involved 30 participants who were asked to evaluate 20 randomly generated results.

As demonstrated in Tab. 1, our proposed method outperforms SOTAs in terms of *SSIM*, *FID*, and *KID*, and achieves a competitive score comparable to SAL-VTON in *LPIPS*. Among GAN-based methods, SAL-VTON integrates additional landmarks to enhance warping, yielding lower LPIPS scores along with notable FID scores. In contrast, we propose a try-on-specific diffusion model, *CTW*, for precise warping, and the subsequent *CTF* incorporates the warped garment to generate results of higher quality compared to the SOTAs. The superior FID and KID scores achieved by our approach substantiate this outcome. Furthermore, the user study results show that our method achieve better try-on accuracy and detail preservation from human perspectives. On the other side, diffusion-based methods (LaDI-VTON, DCI-VTON), exhibit strong generative capabilities. However, they often face difficulties generating fine details, making it challenging to preserve the clothing designs, resulting in worse scores. Our designed *CTF* preserves the generative capabilities of the diffusion model while ensuring the garment details, as reflected in *LPIPS*, *FID*, and *KID*, which we discussed in the last part of this section.

	Method	Paired		Unpaired		Inference time (s)	User study↑
		SSIM↑	LPIPS↓	FID↓	KID↓		
GAN	VITON-HD	0.866	0.134	12.27	0.347	0.37	3.67%
	HR-VITON	0.878	0.115	11.91	0.334	0.85	17.33%
	SAL-VTON	0.893	**0.092**	9.84	0.171	1.87	23.17%
Diffusion	LaDI-VTON	0.867	0.172	11.69	0.412	22	14.83%
	DCI-VTON	0.879	0.160	11.28	0.360	53	17.17%
	Ours	**0.899**	0.099	**9.79**	**0.109**	1.26	**32.50%**

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

Table 1. Quantitative comparison for try-on results.

Inference time comparison. In Tab. 1, LA-VTON shows 17x faster compared to SOTA diffusion-based methods and similar inference times to GAN-based methods. This can be attributed to the utilization of the diffusion process in the latent space and the fewer cross-attention layers in our model design, which significantly enhance our efficiency.

Comparative analysis of artifact reduction using FID and KID. With the most significant cases between LaDI-VTON and ours in Tab. 1, our method has 16.2% improvement in FID but 73.5% improvement in KID. The large difference between FID and KID improvement is critical evidence of our low-artifact performance. Firstly, FID is calculated based on the Gaussian distribution, giving high scores when the generated results generally follow the distribution. On the contrary, since KID is a non-parametric test, it tends to be more robust and sensitive to detailed improvement, e.g., artifacts, complex clothing patterns. In conclusion, the marked disparity in improvements between FID and KID not only underscores the effectiveness of our method in reducing artifacts but also affirms the robustness and precision in complex image generation tasks. *Please refer to supplements for additional experiments and vision results.*

4. Conclusion

In this paper, we propose LA-VTON, addressing key challenges in visual quality: (i) clothing warping stability and (ii) texture fusing with consistency. LA-VTON contains two diffusion-based modules: (i) *Conditional Texture Warping* and (ii) *Conditional Texture Fusing* modules, where we redesign the LDM to reduce visual artifacts and achieve 17× faster inference time. Extensive experiments reveal that LA-VTON outperforms existing SOTAs and delivers remarkable visual enhancements.

Acknowledgment

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

Supplementary Material

5. 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.

6. Additional Ablation Study

6.1. 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 [13]. 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. 6 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. 6.1 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).

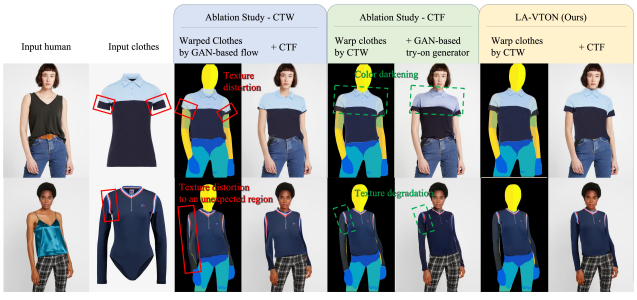

Figure 6. Ablation study of CTW and CTF.

6.2. 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. 7, 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.

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

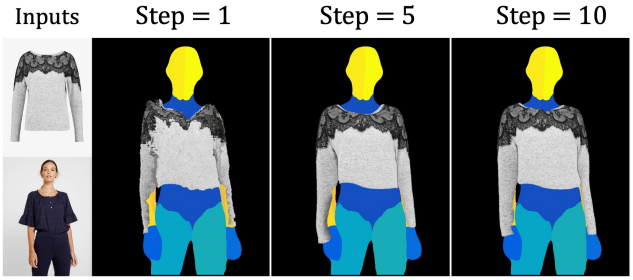

Figure 7. Comparison of different sampling steps.

Figure 8. **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.

6.3. 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. 8 and Sec. 6.3. 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 [14]. 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 $4\times$ lower than our full model prediction. While it worked adequately, it led to obvious misalignment on arms, as shown in Fig. 8. Moreover, we experimented with different loss functions to assess their impact on clothing alignment, as illustrated in Fig. 8. In addition, Sec. 6.3 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 3. Quantitative comparison for different training objectives.

6.4. 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. 6.4, while the qualitative outcomes are presented in Fig. 9. 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 4. Quantitative comparison for different loss weights.

6.5. 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

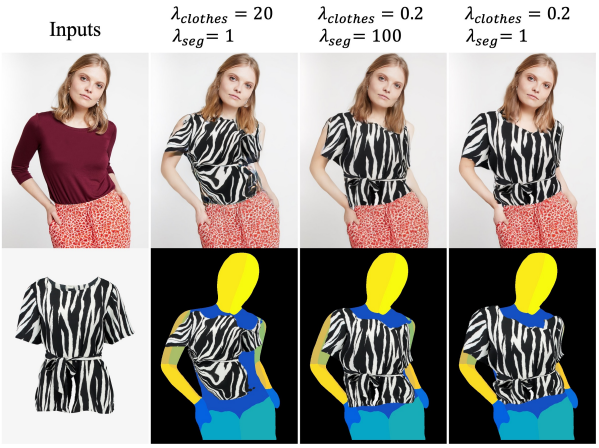

Figure 9. Comparison of different λ settings in CTW module.

of C_{cond} is derived as follows:

$$Train : C_{cond} = I_c \odot \mathcal{W}(C_{mask}, \hat{x_0}), \quad (5)$$

$$Test : C_{cond} = C_{warp} \odot \hat{S}_c, \quad (6)$$

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

$$Train : C_{cond} = I_c, \quad (7)$$

$$Test : C_{cond} = C_{warp}. \quad (8)$$

As shown in Fig. 10, 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.

7. Occlusion Results

Fig. 11 illustrates cases of occlusion, demonstrating that our model is capable of generating good results even when the arms obstruct the clothing.

8. 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. 12 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.

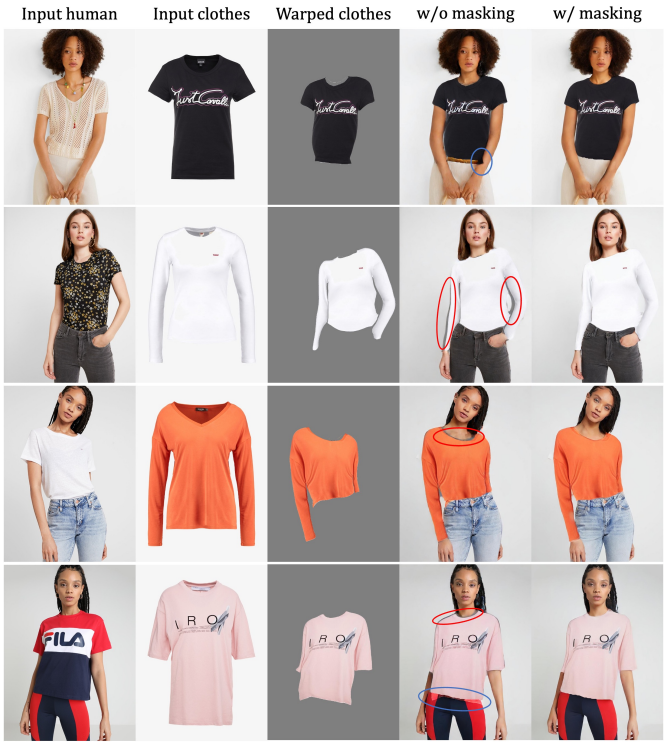

Figure 10. **Comparison of different C_{cond} in CTF module.** The column of w/o masking is the results of using only I_c and C_{warp} as C_{cond} in training and inference time.

Figure 11. Occlusion cases.

Figure 12. Failure cases of complex poses.

Incorrect Mask. As Fig. 13 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.

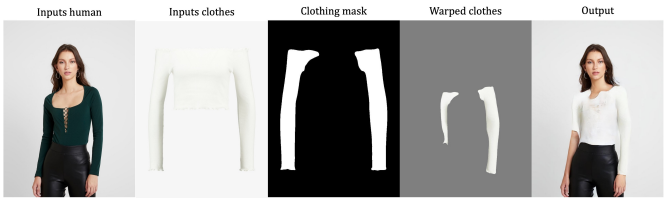

Figure 13. Failure cases of incorrect masks.

9. Additional Qualitative results

We provide additional qualitative comparisons in Figs. 14 to 19. Fig. 14 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. 15 to 17 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. 15 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. 16 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. 17 demonstrates the performance on preserving patterns around the neckline and bottom of clothes. Additionally, Figs. 18 and 19 showcase the outperforming texture-preserving capabilities of our method.

References

[1] Chieh-Yun Chen, Ling Lo, Pin-Jui Huang, Hong-Han Shuai, and Wen-Huang Cheng. Fashionmirror: Co-attention feature-remapping virtual try-on with sequential template poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[2] Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[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 (CVPR)*, 2021. 1, 2, 3

[4] Chien-Lung Chou, Chieh-Yun Chen, Chia-Wei Hsieh, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. Template-free try-on image synthesis via semantic-guided optimization. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021.

[5] Ruili Feng, Cheng Ma, Chengji Shen, Xin Gao, Zhenjiang Liu, Xiaobo Li, Kairi Ou, Deli Zhao, and Zheng-Jun Zha. Weakly supervised high-fidelity clothing model generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[7] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[8] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[9] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2023. 1, 2, 3

[10] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[11] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, and Wen-Huang Cheng. Fit-me: Image-based virtual try-on with arbitrary poses. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019. 1

[12] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2019.

[13] 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, 3

[14] 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. 3, 2

[15] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C. Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[16] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual

Try-On. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2023. 1, 2, 3

[17] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 1

[18] 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 (CVPR)*, 2022. 1, 3

[19] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[20] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[21] Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[22] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] Xie Zhenyu, Huang Zaiyu, Dong Xin, Zhao Fuwei, Dong Haoye, Zhang Xijin, Zhu Feida, and Liang Xiaodan. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[24] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

Figure 14. 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.

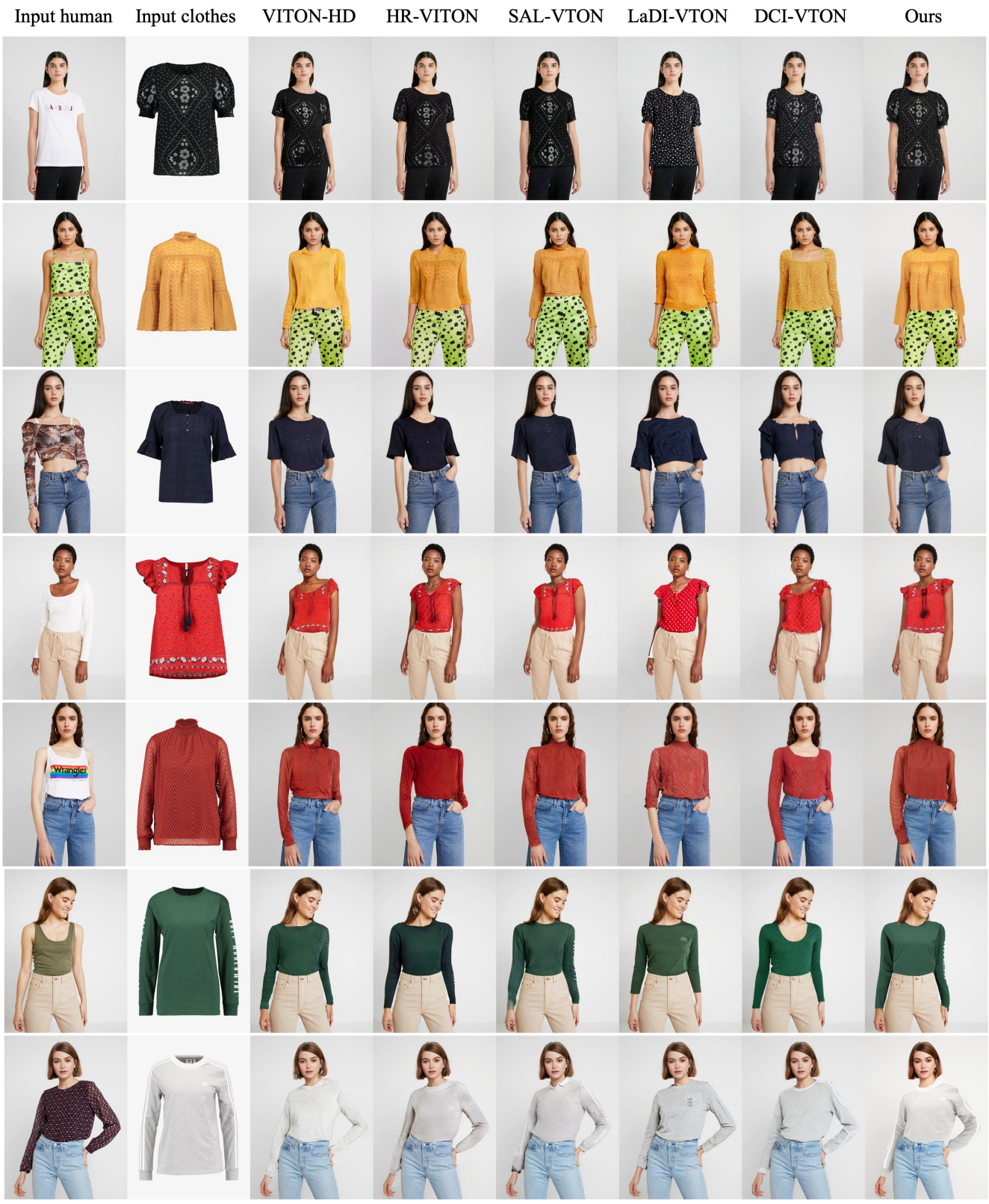

Figure 15. Additional comparison with state-of-the-art try-on methods. It highlights our ability to preserve the special shape of sleeves, *e.g.*, puff sleeves, shoulder pad on T-shirt, and text/line design on side arm.

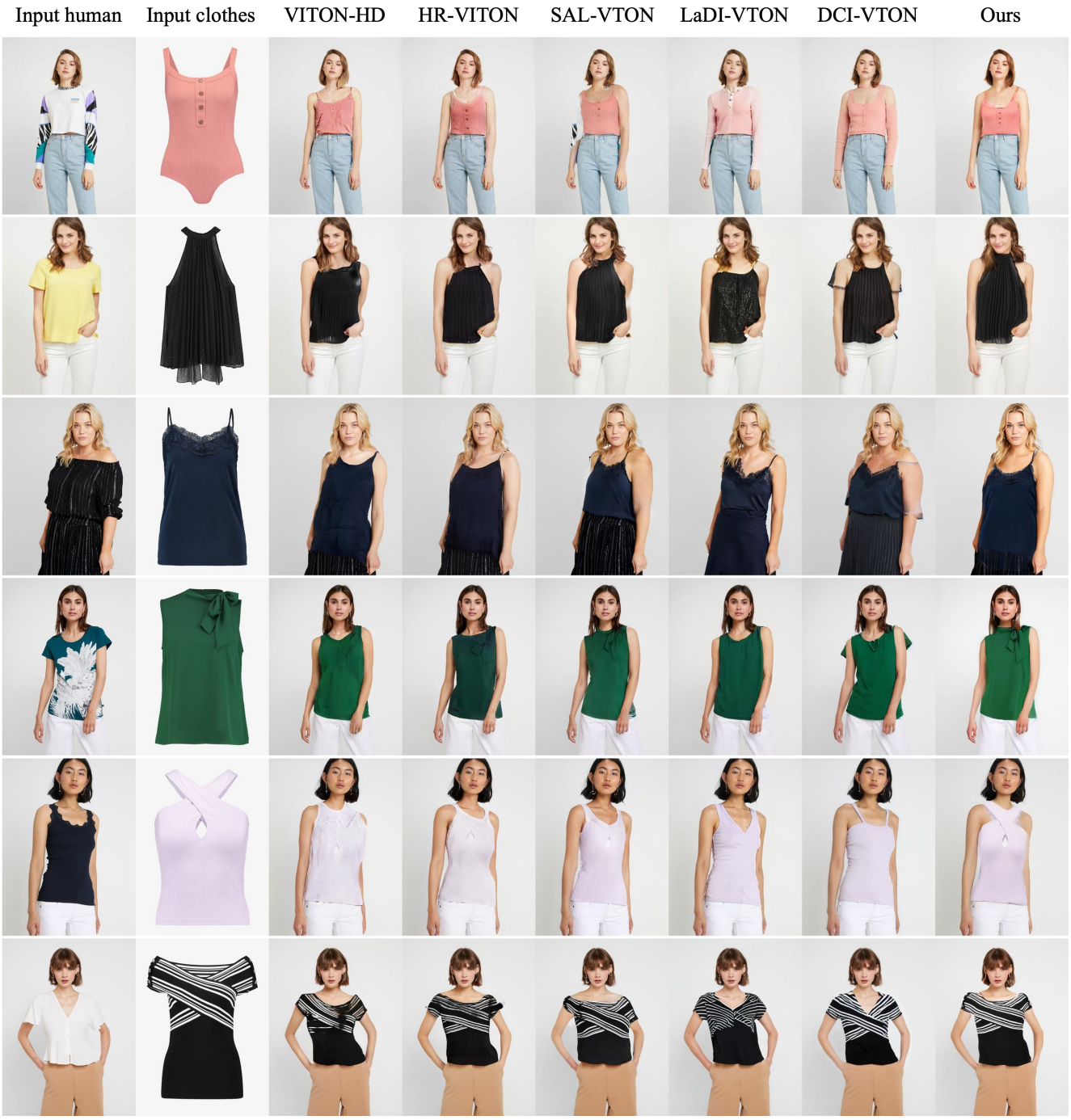

Figure 16. Additional comparison with state-of-the-art try-on methods. Our method accurately generates try-on results for side bow tie designs, cross-strap vests, which are rare clothing styles in the VITON-HD dataset.

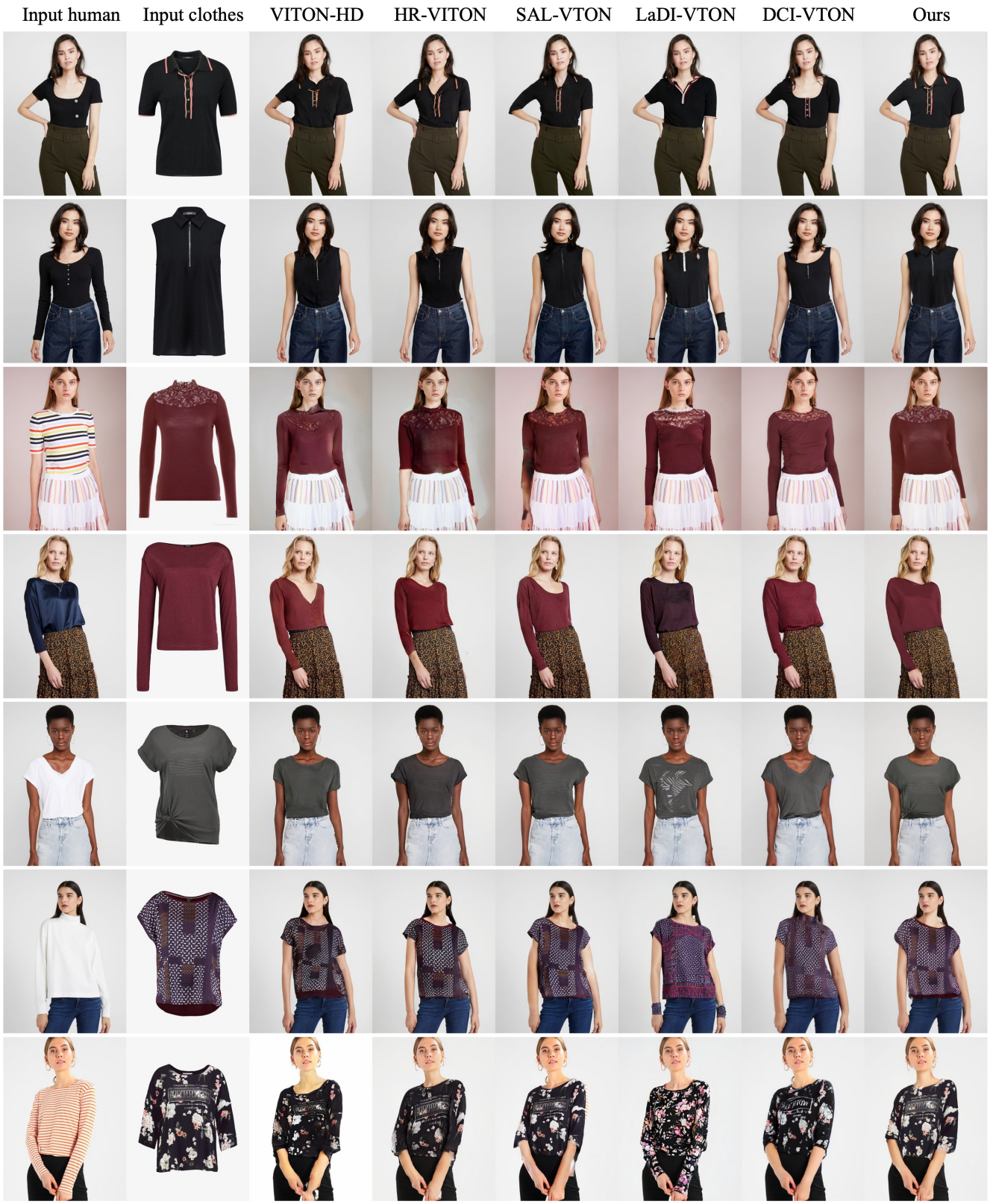

Figure 17. Additional comparison with state-of-the-art try-on methods. It demonstrates the performance on preserving patterns around the neckline and bottom of clothes.

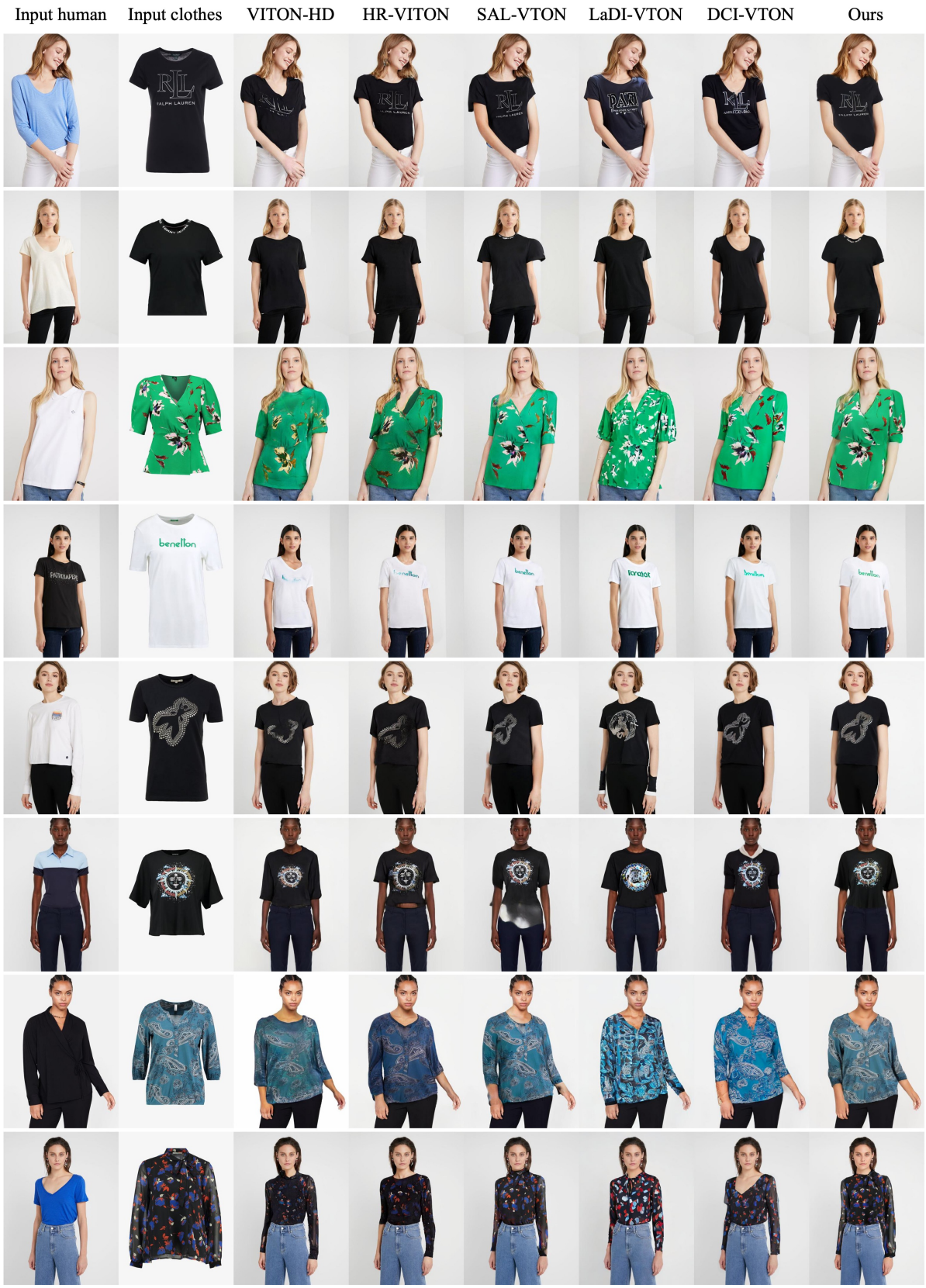

Figure 18. Additional comparison with state-of-the-art try-on methods. Our method showcase the outperforming texture-preserving capabilities.

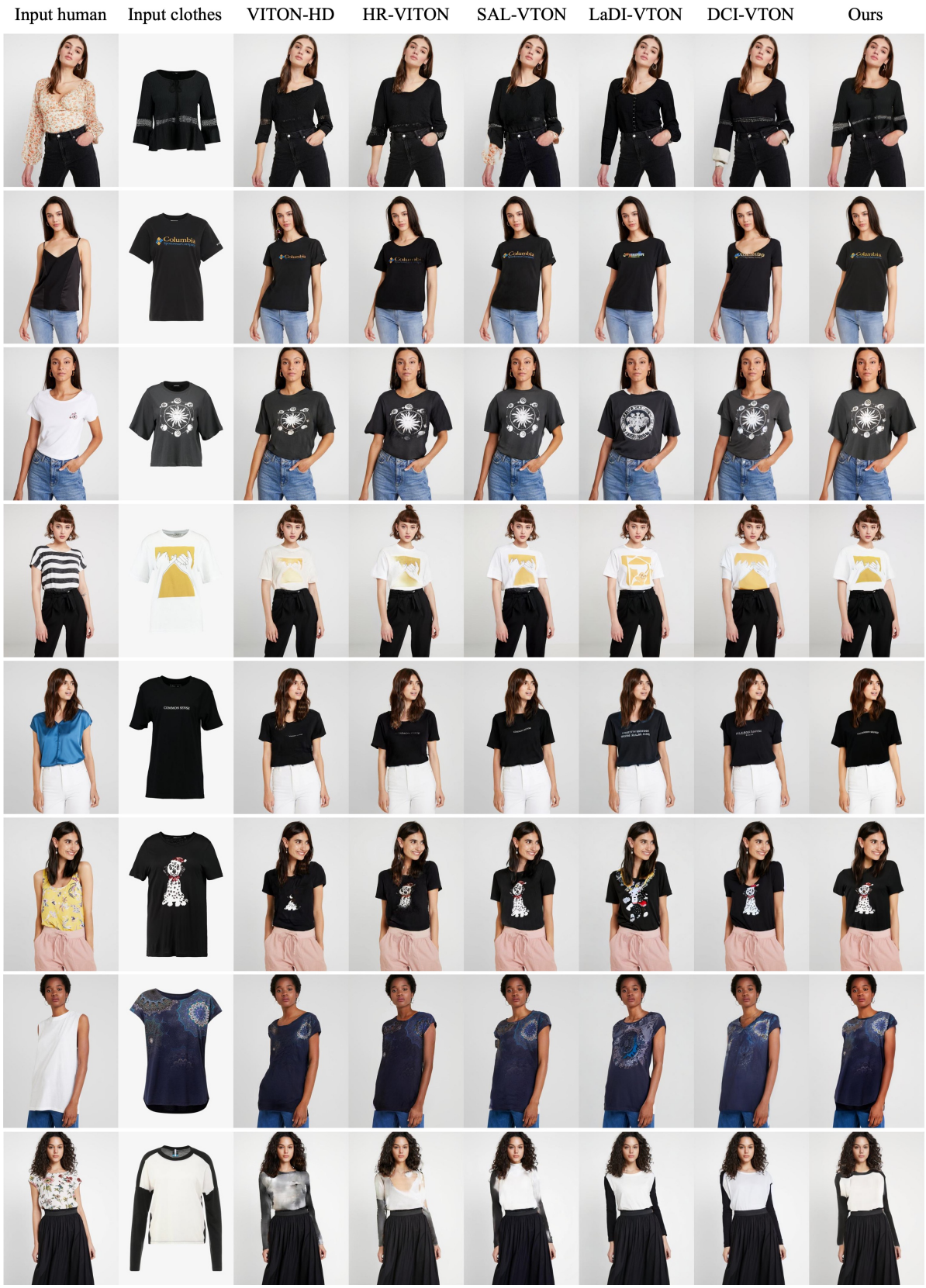

Figure 19. Additional comparison with state-of-the-art try-on methods. Our method showcase the outperforming texture-preserving capabilities.