PICTURE: PhotorealistIC virtual Try-on from UnconstRained dEsigns
Supplementary Material

1. Preliminaries: Stable Diffusion

As a SOTA diffusion model, Stable Diffusion [7] consists of an autoencoder $\mathcal{A}$ containing encoder $\mathcal{E}$ and decoder $\mathcal{D}$, a U-Net ϵ_θ with trainable parameter θ, and a CLIP encoder $\mathcal{T}$. During training, the encoder $\mathcal{E}$ maps a training image $I \in \mathbb{R}^{H\times W\times 3}$ from pixel space to the latent space in $\mathbb{R}^{h\times w\times 4}$, where $h = \frac{H}{8}$ and $w = \frac{W}{8}$. Then, ϵ_θ is trained using the following loss function:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(I),\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2], \quad (1)$$

where t represents the time step, $c = \mathcal{T}(Y)$ is the condition extracted from caption Y, z_t is the latent $\mathcal{E}(I)$ with Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ added stochastically. During inference, a latent is sampled from Gaussian noise, which is then denoised by ϵ_θ under the guidance of c for T time steps. Finally, the decoder $\mathcal{D}$ maps the denoised latent back to the pixel space to get the generated image.

2. More Details about Datasets

We conducted experiments on the DeepFashion Multimodal [6], SHHQ [3] and VITON-HD [2] datasets:

- DeepFashion-Multimodal contains 12,701 full-body images and their corresponding text descriptions. Following FashionTex [5], we randomly divided the images into 11,265 for the training set and 1,136 for the testing set.
- SHHQ is composed of 40,000 full-body images, with the first 35,000 images used as the training set, and the remaining as the testing set.
- VITON-HD comprises 11,647 images for the training set and 2,032 images for the testing set.

3. More Details about Feature Clustering

As mentioned in Sec. 4.2 of the main paper, we cluster the CLIP features into eight categories to balance their contributions. For this purpose, we compute the average cluster groups over 40,000 pieces of clothing and apply the same grouping strategy to all samples, which is not only effective in identifying representative features but also computationally efficient.

4. More Details about User Study

To complement Tables 2 and 4 of the main paper, we use a different method for comparison which assigns numerical scores to the rankings as follows: For n methods, the 1st rank gets n points, 2nd rank gets $n-1$ points, and 3rd rank gets $n-2$ points, etc. As shown in Table 1, our method still receives the highest average score compared to SOTA ones, indicating that our method is the most favorable by users.

5. Additional Qualitative Results

ucVTON. Figs. 1, 2, 3 and 4 show additional qualitative results of our method using different style and texture conditions, which further demonstrates its robustness and generalizability. Moreover, Figs. 1 and 2 use the same style inputs, but with the texture inputs in reversed order. This demonstrates that our method successfully disentangles the texture from the input texture images without being affected by their styles.

Demo. We include a demo in the supplementary materials to explain our idea in a more intuitive way.

Texture Transfer. In Fig. 5, we show additional results on garment texture transfer, which further demonstrates that our method achieves excellent texture transfer results on a wide range of texture patterns (*e.g.*, pure color patterns, floral patterns, stripe patterns and plaid patterns).

In-shop Virtual Try-on. As Fig. 6 shows, our method achieves comparable results to SOTA ones.

6. Limitations and Future Work

For future work, we plan to explore adding user controls over garment shape and fit to further improve the virtual try-on experience. We hope our work will inspire more research into unconstrained VTON to enable highly customizable and personalized outcomes.

Figure 1. More results of our ucVTON. Red text: style for upper part; Blue text: style for lower paer; Red box: texture for upper part; Blue box: texture for lower part.

Figure 2. More results of our ucVTON. Red text: style for upper part; Blue text: style for lower paer; Red box: texture for upper part; Blue box: texture for lower part.

Figure 3. More results of our ucVTON. Red text: style for upper part; Blue text: style for lower paer; Red box: texture for upper part; Blue box: texture for lower part.

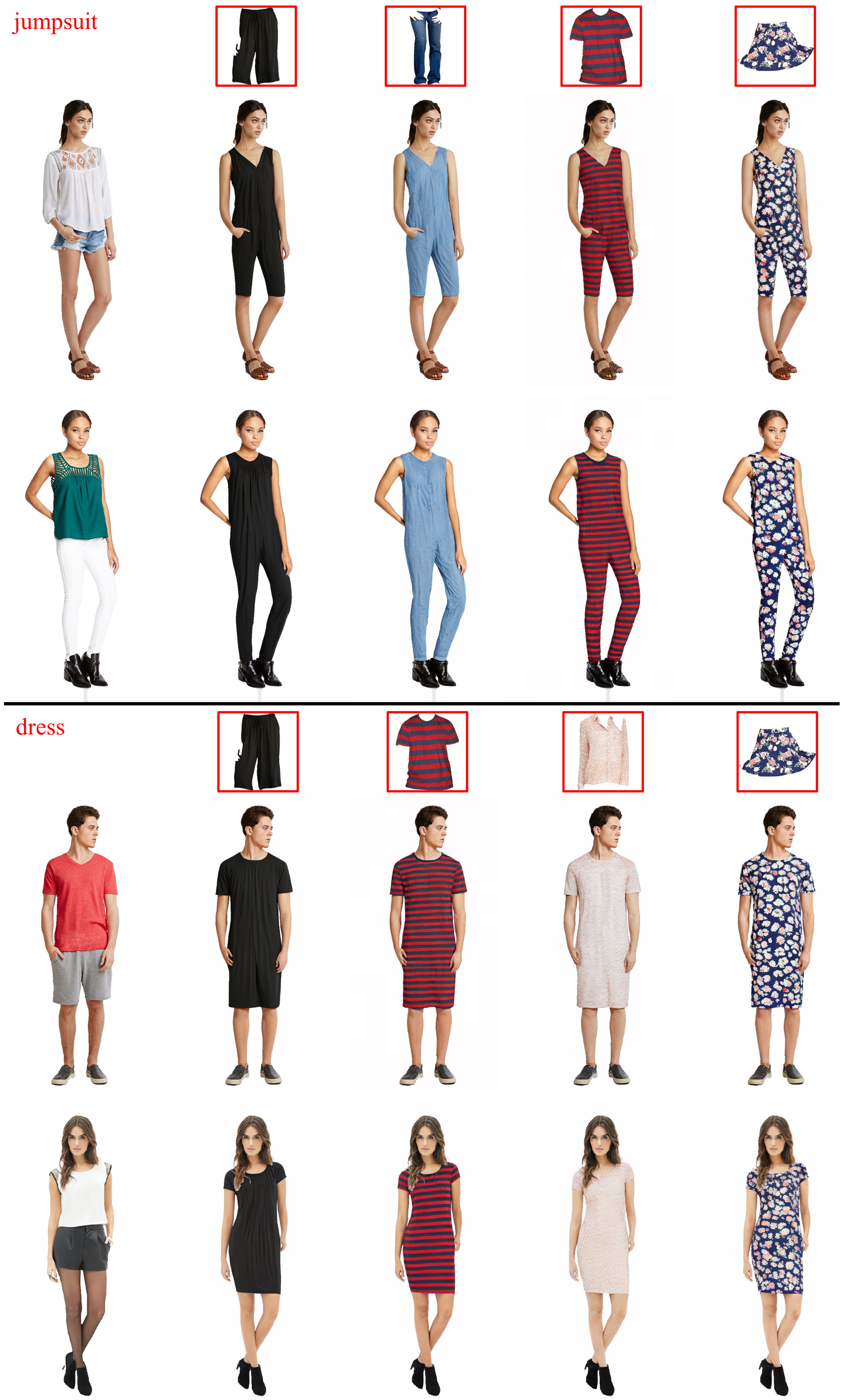

Figure 4. More results of our ucVTON. Red text: style reference; Red box: texture reference.

Figure 5. Visual comparison on garment texture transfer.

Methods	Style		Texture patch		Garment	
	M ↑	R ↑	M ↑	R ↑	M ↑	R ↑
Texture Reformer[8]	–	–	2.58	1.64	1.77	0.93
Paint-by-Example [9]	–	–	1.88	2.41	1.10	1.76
PIDM[1]	–	–	–	–	2.56	1.98
FashionTex [5]	1.24	1.66	1.72	2.14	–	–
Text2Human [4]	1.89	1.59	–	–	–	–
Ours	**2.86**	**2.77**	**3.81**	**3.82**	**3.52**	**3.53**

Table 1. Additional user studies to objectively compare our methods with others at style fidelity, texture fidelity and image naturalness.

References

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023. 6

[2] 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. 1

[3] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *arXiv preprint*, arXiv:2204.11823, 2022. 1

[4] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 6

[5] Anran Lin, Nanxuan Zhao, Shuliang Ning, Yuda Qiu, Baoyuan Wang, and Xiaoguang Han. Fashiontex: Controllable virtual try-on with text and texture. *arXiv preprint arXiv:2305.04451*, 2023. 1, 6

[6] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[7] 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

[8] Zhizhong Wang, Lei Zhao, Haibo Chen, Ailin Li, Zhiwen Zuo, Wei Xing, and Dongming Lu. Texture reformer: towards fast and universal interactive texture transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2624–2632, 2022. 6

[9] 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. 6

Source Garment DCI-VTON Ladi-VTON Ours

Figure 6. The comparison of virtual try-on based on in-shop cloth.