UniHuman: A Unified Model For Editing Human Images in the Wild

Supplementary Material

In the following, we first discuss related literature in Sec. 1. Then we analyze our text manipulation results in Sec. 2 and show more ablation experiments in Sec. 3. Subsequently, we explain how to incorporate LH-400K dataset into training in Sec. 4 and show that our unified model can achieve multi-task combinations in Sec. 5. Finally, we explain our implementation details in Sec. 6 and present various visualized examples in Sec. 8.

1. Task-Specific Models Discussion

In the following, we discuss existing task-specific models that achieve reposing, virtual try-on, and text manipulation in the 2D image domain and how they differ from our proposed approach. Since our goal is human *image* editing, we do not compare with 3D and video based models [8, 22, 30, 33, 38] as they are out-of-scope of this paper.

Reposing. To change a person's pose, previous approaches typically encode the person as a whole and learn the transformation across poses [2, 4, 18, 28, 32, 42]. Adapting these models to handle multi-task scenarios where only specific body parts need to be modified can be difficult. Other methods enhance the versatility of reposing models by dissecting the person into different parts [4, 7, 20, 23, 34], thus allowing independent editing of each part. However, relying solely on part-wise texture information may lead to challenges in recognizing the identities of individual body parts. For example, a strapless top might be misidentified as a mini skirt since both clothing could have similar textures and shapes. To address this concern, a word embedding labeling the clothing type, such as *upper clothing* and *lower clothing*, is concatenated with the DINOv2 features. Furthermore, we apply a loss L_B in **??** to localize the cross-attention map of each human part to its corresponding region. We find these strategies effectively improve the performance of our model.

For reposing methods that introduce a pose-warping module [1, 13], a UV coordinate inpainting model was trained to infer invisible pixels from their visible counterparts, which is unsuitable for warping in-shop garments that lack such UV representation. As a unified model, UniHuman can utilize both dense pose UV representation and sparse keypoint locations to warp clothing texture to the RGB space, ensuring the provision of accurate visible pixel information across domains.

Virtual Try-on. The objective of virtual try-on is to seamlessly fit the target in-shop clothing to a person [5, 19, 26, 36, 43]. In prior work, this is often accomplished through a two-stage process where the clothing is initially warped through a deep learning model and subsequently aligned with the person in a second model [5, 8, 19, 21, 37, 40]. The clothes warping module often learns the parameters of a Thin-Plate Spline transformation (TPS) [9] from the target garment to the target pose [40]. To balance the flexibility of TPS with the rigidity of affine transformation, researchers have introduced various regularization terms to train these parameters [11, 40, 41]. However, these learned warping modules are trained using try-on data to establish correspondences between the clothing and the pose, posing a risk of overfitting to specific body shapes within the dataset. In contrast, our pose-warping module harnesses the pose correspondences to map the original pixels to pose-warped texture without training. Moreover, while TPS is not suitable for the pose transfer task due to the non-smoothness of the pose transformation, our pose-warping model can leverage dense pose for texture warping.

Diffusion Based Text Manipulation. The advent of diffusion models has ushered in a new era in image editing through text descriptions [3, 12, 15, 27, 39]. Among these methods, latent Stable Diffusion (SD) [29] has gained popularity because of its versatility in accommodating prompts of various formats, coupled with its efficient memory utilization within the latent space. However, the challenge of editing human images using text prompts persists, primarily due to the highly structured nature of the human body [4, 10, 17]. Additionally, enhancing the alignment between text and images requires the image captions to include information about clothing categories, shapes, and textures. In our pursuit of expanding existing human image-text datasets [17], we curated a new dataset featuring single human images paired with captions from LAION-400M [31]. We believe that incorporating these image-text pairs in the human image editing tasks can further improve the data diversity and enrich the modality of our model.

2. Text Manipulation Analysis

Results. We randomly chose 1000 images from the test sets for text manipulation evaluation. Tab. 1 reports the FID, KID, and CLIP image-text similarity scores for all the comparison methods. UPGPT [4] is a multi-task model. EditAnything [10] is an SD-based text manipulation model. Our UniHuman shows better performance on all these metrics, demonstrating its capacity for human-specific text manipulation. For further evaluation, we also conducted a user study. We asked AMTurk workers to compare 200 samples edited by our model and by an existing method on WVTON. The workers evaluate the image quality on three

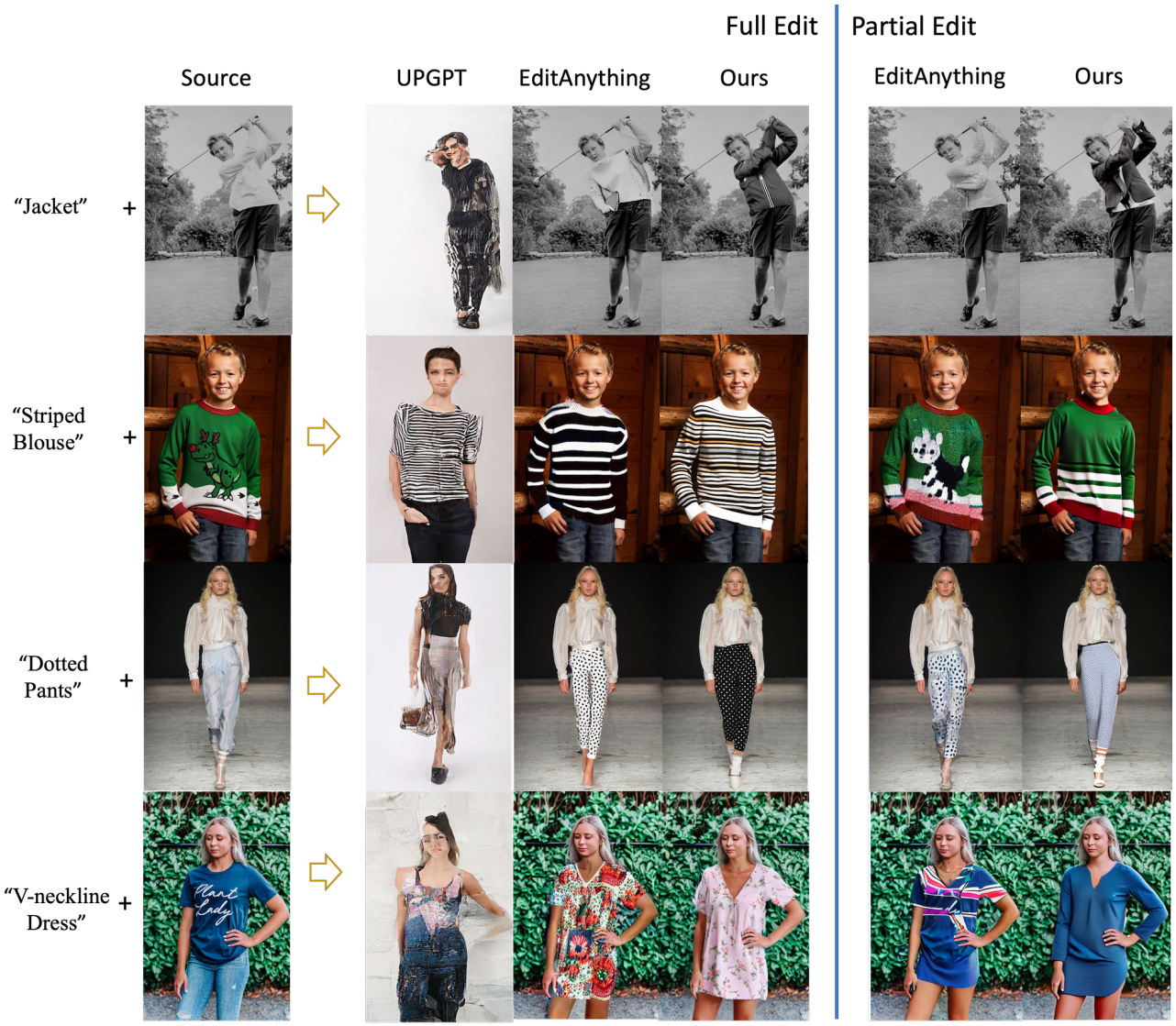

Figure 1. Text manipulation examples. Full Edit indicates a new garment is synthesized from scratch. Partial Edit means partial textures from the source garment are used to generate the new clothing. Zoom in to see details.

	FID↓	KID↓	CLIP↑
UPGPT [4]	138.178	8.548	13.628
EditAnything [10]	63.336	1.986	15.235
UniHuman	**62.663**	**1.782**	**15.715**

Table 1. Quantitative results for text manipulation. Our UniHuman shows outperforms the baselines in all metrics.

Methods	EditAnything [10]	UniHuman	UPGPT [4]	UniHuman
Pose Accuracy	32.6%	67.4%	13.2%	86.8%
Image Plausibility	35.8%	64.2%	15.9%	84.1%
Image-Text Similarity	39.5%	60.5%	17.8%	82.2%

Table 2. Human evaluation results on WVTON for text manipulation. Our UniHuman is preferred by users on all three evaluation methods.

aspects: pose accuracy, image plausibility, and image-text similarity. Tab. 2 reports the human evaluation results. Our UniHuman outperforms prior work on all three aspects in maintaining the original pose and manipulating the texture, acknowledged by at least 60% workers.

Visulizations. Fig. 1 presents different ways of editing human images by text descriptions: the left side of the vertical ruler includes examples of generating a new garment from scratch, denoted by Full Edit; the right side of the vertical ruler shows visualizations of editing the garment given random partial source textures, denoted by Partial Edit. We did not compare with UPGPT [4] in Partial Edit since it can not achieve partial editing by design. Compared with EditAnything [10], our UniHuman shows better capability at

		In-Domain							Out-of-Domain				
		DeepFashion [17]			VITON-HD [5]		DressCode [25]		WPose			WVTON	
		FID↓	SSIM↑	LPIPS↓	FID↓	KID↓	FID↓	KID↓	FID↓	M-SSIM↑	M-LPIPS↓	FID↓	KID↓
(c)	pix seg	5.785	0.808	0.126	9.638	0.269	6.345	0.218	30.421	0.808	0.160	134.575	1.964
	w.o. emd	5.456	0.811	0.125	9.562	0.252	6.433	0.209	28.435	0.808	0.161	132.802	1.740
	w.o. L_B	5.827	0.813	0.126	10.086	0.310	6.605	0.239	29.175	0.805	0.162	134.021	1.856
	w.o. pretrain	5.328	0.811	0.128	9.749	0.270	6.664	0.252	31.287	0.805	0.166	135.287	1.998
	UniHuman	**5.089**	**0.815**	**0.123**	**9.558**	**0.248**	**6.310**	**0.208**	**27.571**	**0.810**	**0.159**	**131.500**	**1.730**
(d)	rp only	5.682	0.813	0.124	17.345	1.005	14.909	0.867	42.587	0.790	0.185	170.322	3.180
	vt only	18.081	0.755	0.210	9.721	0.255	7.012	0.280	76.605	0.784	0.208	143.137	2.378
	tm only	9.773	0.751	0.204	20.623	1.429	18.365	1.152	62.219	0.789	0.196	175.602	5.387
	multi-task	**5.269**	**0.826**	**0.110**	**9.659**	**0.184**	**6.960**	**0.271**	**37.826**	**0.804**	**0.167**	**142.839**	**2.139**

Table 3. Ablation results on 256x256 images. KID is multiplied by 100. Our full model achieves the best overall performance.

making the edited image more plausible. For example, on the first row of Full Edit, our generated image better fits the jacket into the person playing golf.

3. Additional Ablations

Part Encoder. Our part encoder includes two modules: a DINOv2 image encoder for visual representation encoding and a CLIP Text encoder for semantic representation encoding. As mentioned in Sec. 1, the CLIP text embedding labeling each body part helps identify these human parts. In Tab. 3(c), **w.o. emd** removes the text embedding, causing a slight drop in all evaluation metrics. In another setting, to prove the effectiveness of the introduced L_B, we removed this loss function in **w.o.** L_B. Results show that the performance on both in-domain and out-of-domain test sets suffers from a larger drop than **w.o. emd**, indicating the importance of L_B in our objective function. Additionally, **pix seg** extracts the part features at a pixel level to compare with our feature-level body part segmentation. The slight drop in all metrics shows that the contextual information from the feature-level segmentation helps reconstruct the clothing textures for all three tasks.

SD Pretraining. We chose SD as our backbone for its excellence in producing text-aligned images and its suitability for multi-task learning. In Tab. 3(c), **w.o. pretrain** model takes 67% more time to converge (5 days vs. 3 days) and shows a slight performance drop in the metrics.

Single-Task Ablations. To investigate if our multi-task objectives reinforce single tasks for each other, we designed four ablation models: **rp only** that takes human images from DeepFashion [17] to do the reposing task, **vt only** that uses human and garment images from try-on datasets [5, 24, 25] to do the try-on task, **tm only** that draws image-text pairs from DeepFashion to do the text manipulation task, and **multi-task** that takes all the above data to achieve three task objectives in the same model. In Tab. 3(d), DeepFashion and WPose are test sets for the reposing task; VITON-HD, DressCode and WVTON are evaluation sets for the try-on task. **Multi-task** effectively learns all three tasks and outperforms single-task models on all metrics. This demonstrates that our multi-task objectives indeed reinforce the two visual tasks (*i.e.*, reposing and virtual try-on) by learning them jointly. Note that **multi-task** is renamed as **w.o. 400K** in **??** of the main paper.

4. Leveraging Unpaired Images In Training

The visual tasks in our model require paired images for training. In reposing, we need image pairs of the same person in different poses. Recent research explores the possibility of using readily accessible unpaired human images for the reposing training [20, 23, 34], which prompts us to explore the acquisition and incorporation of less costly unpaired images. Refer to **??** for details on collecting these data. However, the incorporation of unpaired images introduces a significant challenge. It often leads to the potential issue of pose leakage and pose-texture entanglement [20], particularly when the volume of unpaired images surpasses that of the paired ones. To mitigate this issue, we implement strong data augmentation techniques when obtaining the part features in **??**. Beyond routine operations like image cropping, resizing and flip, we meticulously ensure that each human part is randomly warped into *different orientations*. This strategy compels the model to heavily rely on the target pose to accurately restore the original body orientation, successfully addressing pose leakage and promoting a more robust and effective training process.

5. Multi-Task Combinations in UniHuman

The three tasks in this paper can be combined in arbitrary ways to edit the original human image within 50 denoising time steps. Another less efficient way of accomplishing this goal is to apply these tasks sequentially using the corresponding task-specific models. Fig. 2 shows the results of using these two types of task combinations. Sequen-

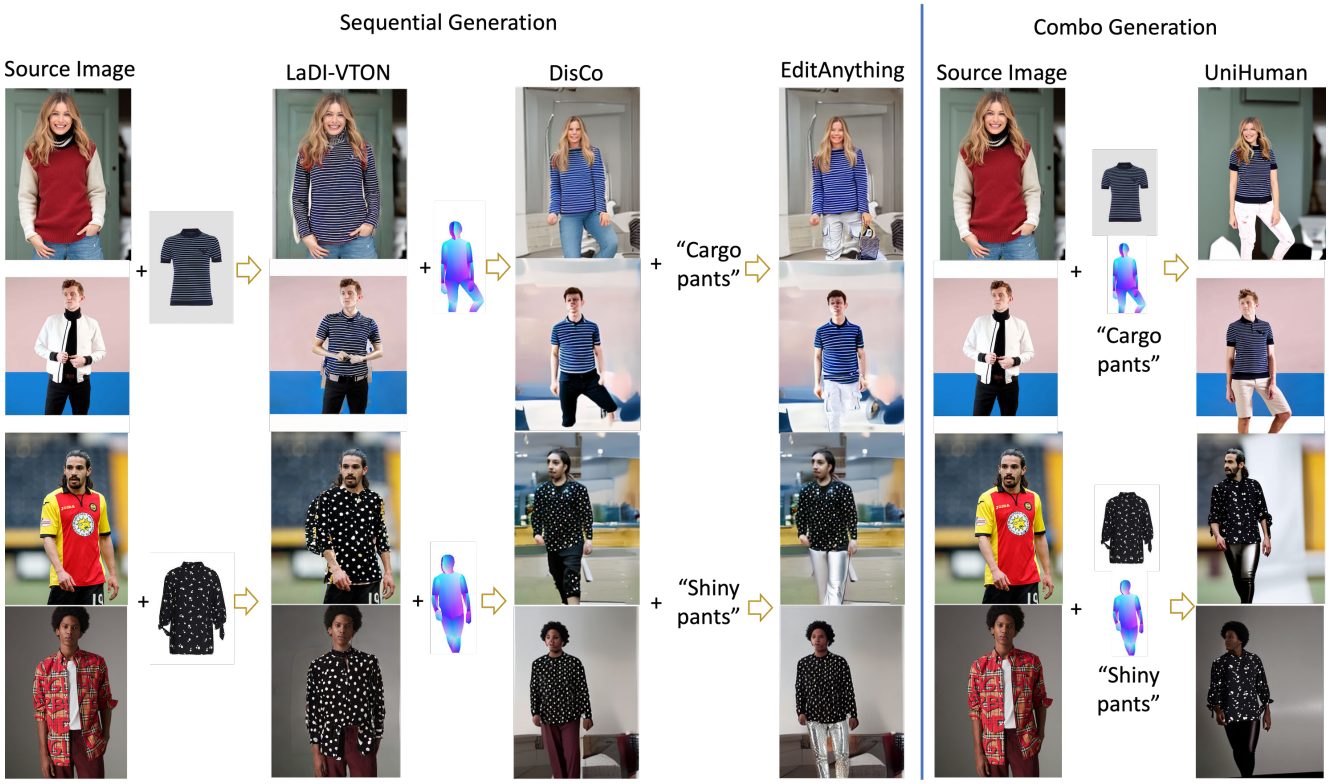

Figure 2. Examples of task combinations. In sequential generation, the input image sequentially goes through the virtual try-on method (LaDI-VTON), the reposing model (DisCo), and the text manipulation approach (EditAnything). In combo generation, our UniHuman can achieve all editing tasks altogether.

tial Generation means applying virtual try-on, reposing, and text manipulation using Ladi-VTON, DisCo and EditAnything, sequentially on the input image. Combo Generation represents achieving all tasks simultaneously in our UniHuman. We find that human images produced by our model better follow the target pose and the given garment textures.

6. Implementation Details

In training, we use pretrained weights from SD v-1.5. IN the part encoder, we use ViT-B/14 for the DINOv2 visual encoding and ViT-B/16 for the CLIP text encoding. Then we fix the SD UNet encoder, the CLIP encoder, and the first 15 blocks of DINOv2, finetuning the rest layers of DINOv2 and SD UNet decoder with a learning rate of 2.5×10^{-5}. The conditioning encoder is trained from scratch with a learning rate of 10^{-4}, which consists of three residual blocks. In our objective function, we set empirically $\lambda_1 = 10^{-3}$ and $\lambda_2 = 2.5 \times 10^{-4}$. Both the 256-resolution model and 512-resolution model are trained for 220K iterations with a batch size of 64. In the dataloader, the target pose, pose-warped texture, part features, and the text prompt have a 10% chance of being zero, respectively. This enables us to use classifier-free guidance [16] when denoising the latent code, improving image quality.

In our part-SD cross-attention, we sequentially apply a global cross-attention using the global feature representation of each human part and a local cross-attention using the patch tokens of each human part. We found this sequential attention to be slightly better than concatenating global features and patch tokens together in one attention block. In the pose-warping module, we run DensePose [14] pretrained on COCO to get the dense pose UV representation and use MMPose [6] pretrained on HKD [35] to obtain the sparse pose keypoint representation.

For inference, we keep the aspect ratio of the original images and resize the longer side to 256/512, then we pad the shorter side to the same size with white pixels. This type of resizing holds the original body shape to be unchanged and keeps as much background as possible. All baseline methods in our experiments use the above resizing for a fair comparison. We set $\epsilon = 2$ to be the guidance scale in the classifier-free guided generation.

In all the tasks, we paste back the background areas that do not change after editing and reconstruct the rest as in the inpainting pipeline. In the reposing task, a large position and pose change could happen, resulting in little background area in the conditioning. On the contrary, in try-

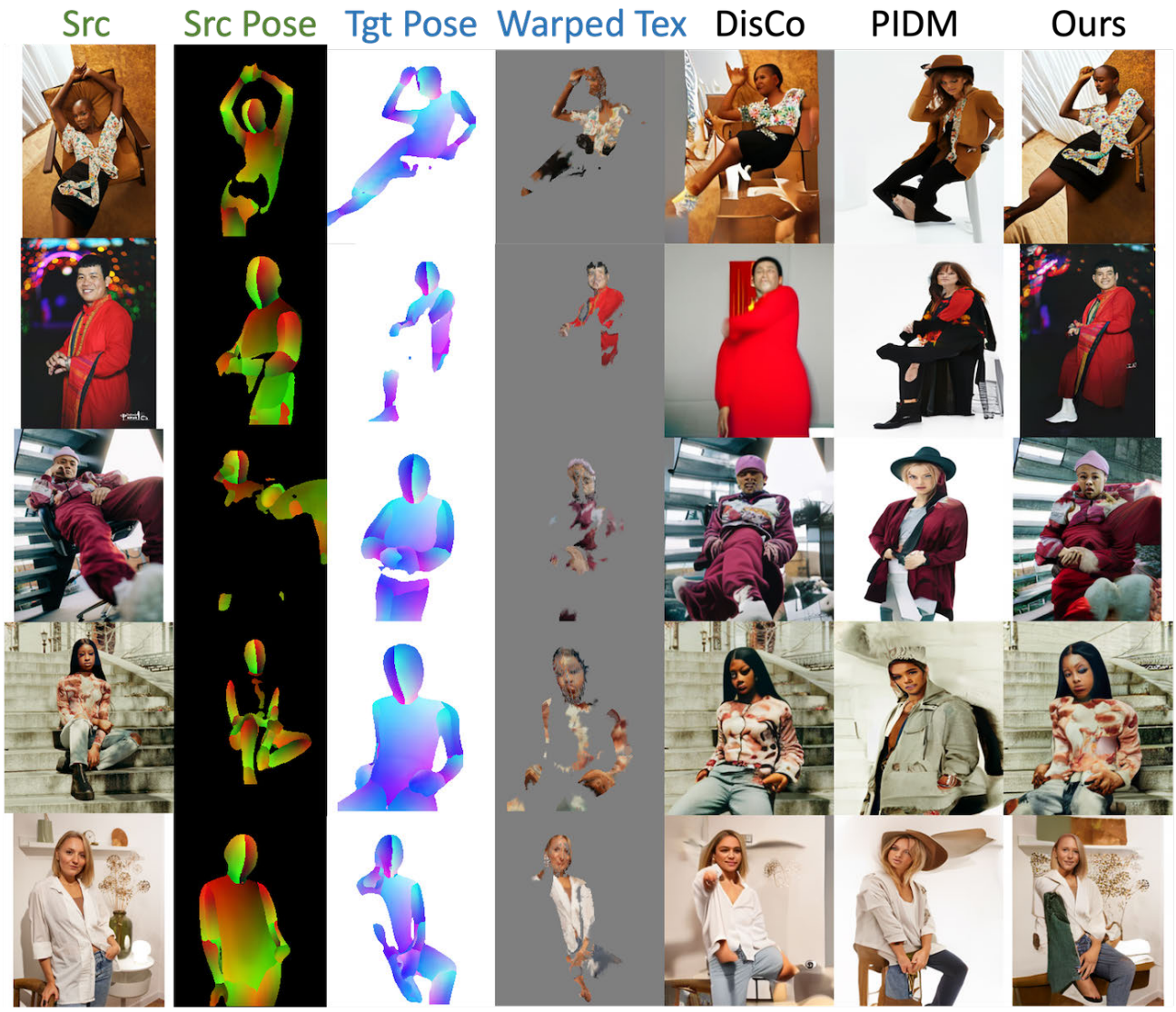

Figure 3. Examples of failed reposing generations. In cases where the source/target densepose predictions are incorrect, our model failed to transfer accurate clothing textures.

on and text manipulation, most background areas do not change and thus require less inpainting.

7. Limitation

Our approach is limited by its reliance on pose detectors and parsing models, which is also a common limitation shared by prior work [2, 4, 32, 37]. This is an even more challenging problem in our WPose dataset, which includes diverse postures with more complicated body part occlusions than standing postures. As a result, the detected densepose could have incorrect predictions and missing parts. In Fig. 3, our model failed to transfer accurate clothing textures due to densepose errors. For future work, we believe incorporating more 3D information, such as depth and surface normal, will help rectify these inaccuracies.

8. Additional Visualized Examples

Fig. 4 and Fig. 6 show in-domain generated results on DeepFashion, DressCode and VITON-HD. Fig. 5 and Fig. 8 give several representative examples from our collected WPose and WVTON, respectively. Results show that images generated by our model are better aligned with the target pose while preserving the face and clothing identities.

Figure 4. Examples of reposing on DeepFashion. Our model better reconstructs the intricate texture patterns.

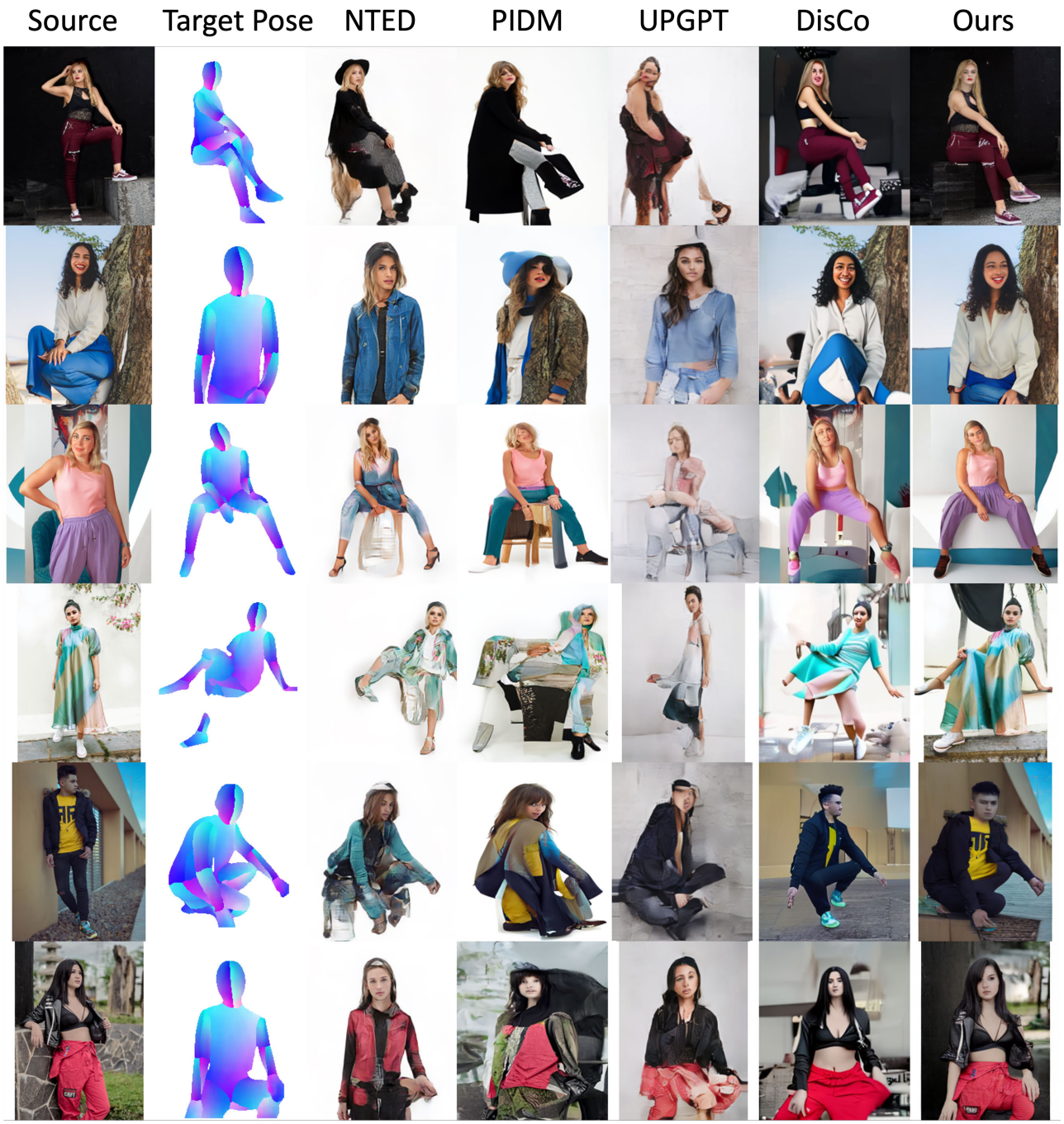

Figure 5. Examples of reposing on WPose. Images generated by our model are better aligned with the target pose while preserving the face and clothing identities.

Figure 6. Examples of virtual try-on on DressCode. Our UniHuman can recover detailed texture patterns.

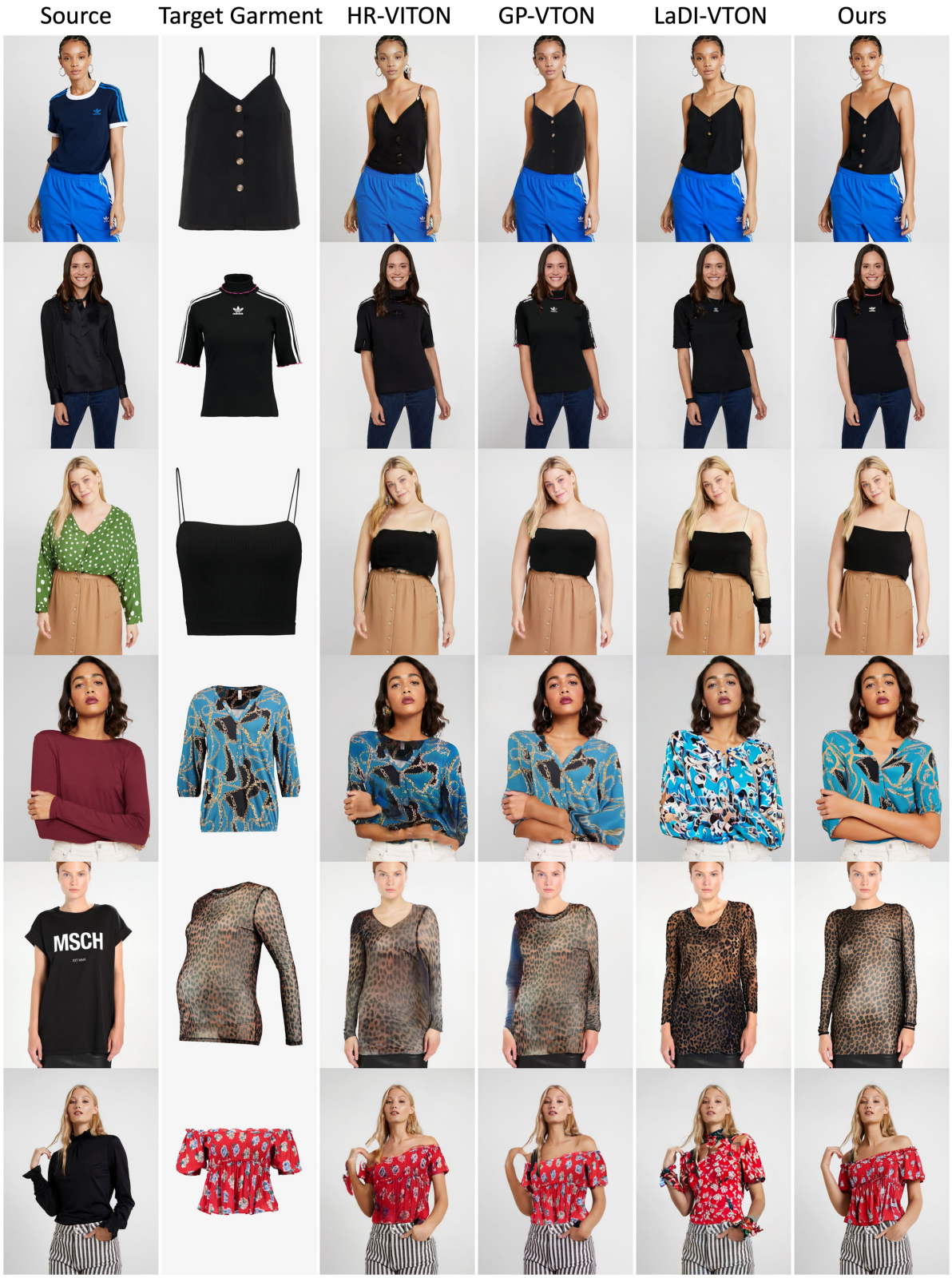

Figure 7. Examples of virtual try-on on VITON-HD. Our model is better at handling occlusions between body parts.

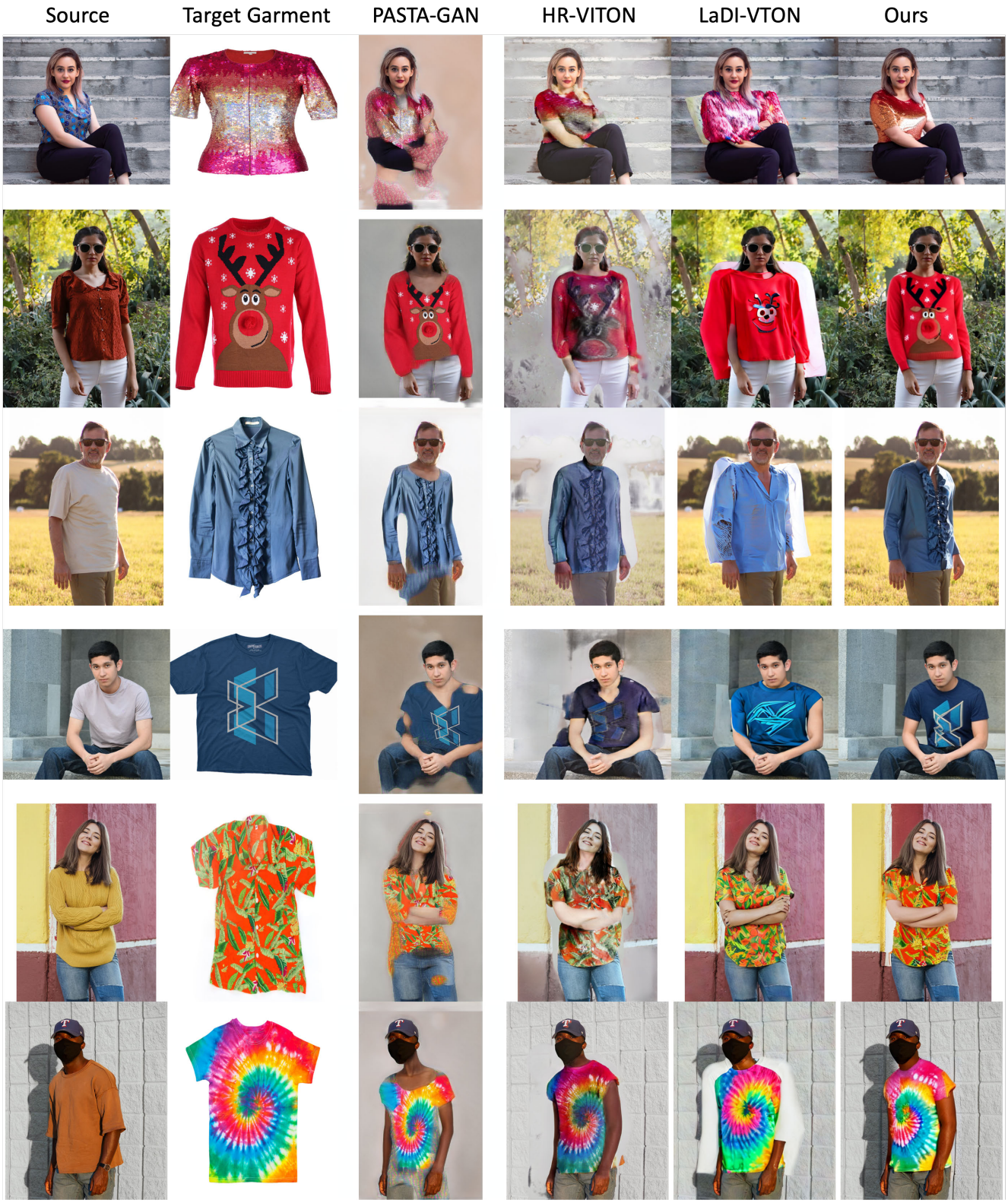

Figure 8. Examples of virtual try-on on WVTON. Our model better fits the new garment onto the person.

References

[1] Badour Albahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with style: Detail-preserving pose-guided image synthesis with conditional styleGAN. *ACM Transactions on Graphics*, 40(6):1–11, 2021. 1

[2] 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*, 2023. 1, 5

[3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of the International Conference on Machine Learning*, 2023. 1

[4] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. UPGPT: Universal diffusion model for person image generation, editing and pose transfer. In *Workshop on the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 5

[5] 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 conference on computer vision and pattern recognition*, 2021. 1, 3

[6] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 4

[7] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1

[8] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[9] Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Proceedings of Constructive Theory of Functions of Several Variables*, pages 85–100, 1977. 1

[10] Shanghua Gao, Zhijie Lin, Xingyu Xie, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. EditAnything: Empowering unparalleled flexibility in image editing and generation. In *Proceedings of the ACM International Conference on Multimedia, Demo track*, 2023. 1, 2

[11] 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*, 2021. 1

[12] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023. 1

[13] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2019. 1

[14] 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*, 2018. 4

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *Proceedings of the International Conference on Learning Representations*, 2022. 1

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[17] 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. 1, 3

[18] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. DreamPose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22680–22690, 2023. 1

[19] 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*, 2022. 1

[20] Nannan Li, Kevin J Shih, and Bryan A Plummer. Collecting the puzzle pieces: Disentangled self-driven human pose transfer by permuting textures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 3

[21] 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*, 2023. 1

[22] Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping GAN with attention: A unified framework for human image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 5114–5132, 2022. 1

[23] Tianxiang Ma, Bo Peng, Wei Wang, and Jing Dong. MUST-GAN: Multi-level statistics transfer for self-driven person image generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021. 1, 3

[24] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 3

[25] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[26] 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*, 2023. 1

[27] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the International Conference on Machine Learning*, 2022. 1

[28] Yurui Ren, Yubo Wu, Thomas H Li, Shan Liu, and Ge Li. Combining attention with flow for person image synthesis. In *Proceedings of the ACM International Conference on Multimedia*, 2021. 1

[29] 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*, 2022. 1

[30] Igor Santesteban, Miguel Otaduy, Nils Thuerey, and Dan Casas. Ulnef: Untangled layered neural fields for mix-and-match virtual try-on. *Advances in Neural Information Processing Systems*, 2022. 1

[31] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[32] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. DisCo: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. 1, 5

[33] Tuanfeng Y. Wang, Duygu Ceylan, Krishna Kumar Singh, and Niloy J. Mitra. Dance in the wild: Monocular human animation with neural dynamic appearance synthesis. In *International Conference on 3D Vision*, 2021. 1

[34] Zijian Wang, Xingqun Qi, Kun Yuan, and Muyi Sun. Self-supervised correlation mining network for person image generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022. 1, 3

[35] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. AI challenger: A large-scale dataset for going deeper in image understanding. In *Proceedings of the IEEE/CVF International Conference on Multimedia and Expo*, 2019. 4

[36] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive gan. In *Advances in Neural Information Processing Systems*, 2021. 1

[37] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 5

[38] Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[39] 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*, 2023. 1

[40] 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 conference on Computer Vision and Pattern Recognition*, 2020. 1

[41] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[42] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022. 1

[43] 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*, 2023. 1