Unified Diffusion Transformer for Bidirectional Virtual Try-On and Try-Off

Seungyong Lee Jeong-gi Kwak
NXN Labs
{seungyong, jeonggi}@nxn.ai

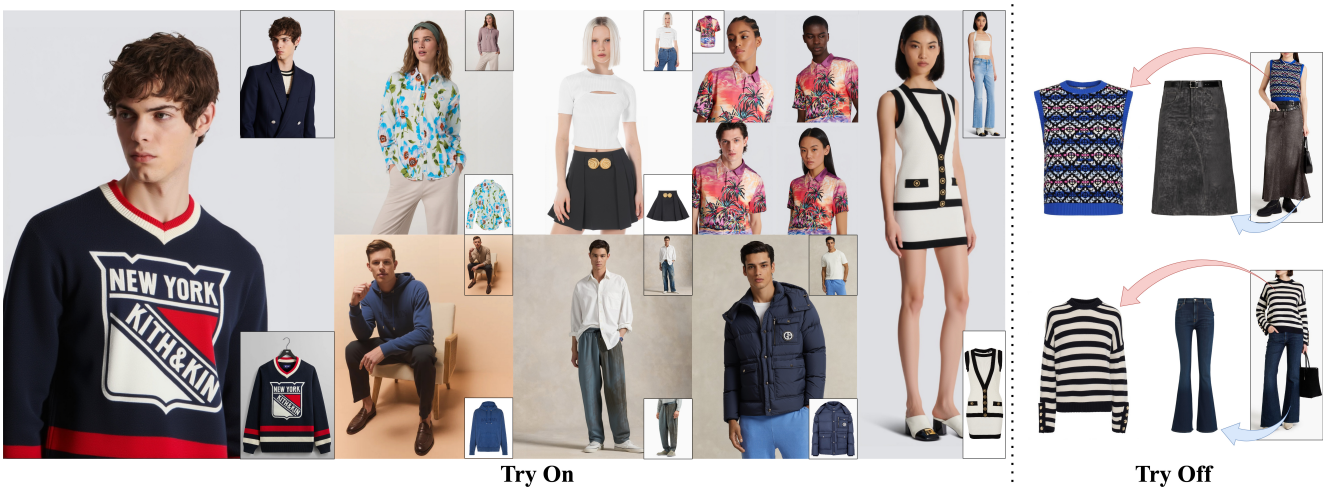

Figure 1. **Teaser –** our model jointly handles virtual try-on and try-off with a single architecture. It achieves high-quality results while maintaining robustness to various human poses, garment categories, and image layouts.

Abstract

We propose a unified framework capable of performing both virtual try-on and virtual try-off within a single diffusion transformer. By jointly learning these two complementary tasks, our approach enhances the garment–target correspondence—a key challenge in virtual try-on. This is achieved through a carefully designed token concatenation structure, which allows spatially aligned garment and person images to be encoded and reasoned about jointly. Remarkably, our unified model achieves state-of-the-art performance on both VTON and VTOFF benchmarks, surpassing prior methods that were specialized for each task.

1. Introduction & Related work

Virtual try-on (VTON) is a generative task that aims to realistically transfer a given garment onto a person image. Early methods were primarily based on GAN-based approaches [1, 4, 8, 16, 19, 28, 30, 33], making notable progress in generating plausible try-on results. However, these models often struggled to preserve fine garment details and to produce high-quality, photorealistic images.

Since then, a number of studies [2, 5, 6, 9, 14, 21, 31, 34, 35] based on diffusion models [7, 11, 23, 24] have brought a significant leap in the performance of virtual try-on systems, particularly in terms of photorealism and garment fidelity. Among them, several approaches [14, 21, 29, 35] that incorporate an auxiliary reference network and leverage its intermediate features through mutual self-attention mechanisms. Although they have shown superior garment preservation, these approaches introduce computational and memory overhead and the domain gap between noisy latent and reference feature leads to suboptimal results.

More recently, a few studies [6, 13] have proposed a simple yet effective approach that spatially concatenates the conditioning image with the input, allowing the model to reconstruct both the target and the conditioning image jointly. As this is effectively equivalent to token concatenation in transformer, it avoids domain gaps and requires no additional networks, offering better memory efficiency. Despite their efficiency, they suffer from a lack of correspondence between the spatially concatenated garment and target images (Fig. 3). As a result, the model often relies on the pretrained diffusion prior rather than faithfully preserving the garment details, leading to reduced fidelity.

To address this issue, several previous works have attempted to enhance alignment by introducing alignment-specific loss functions [14, 34]. However, they often lead to degraded image quality or unnatural synthesis results.

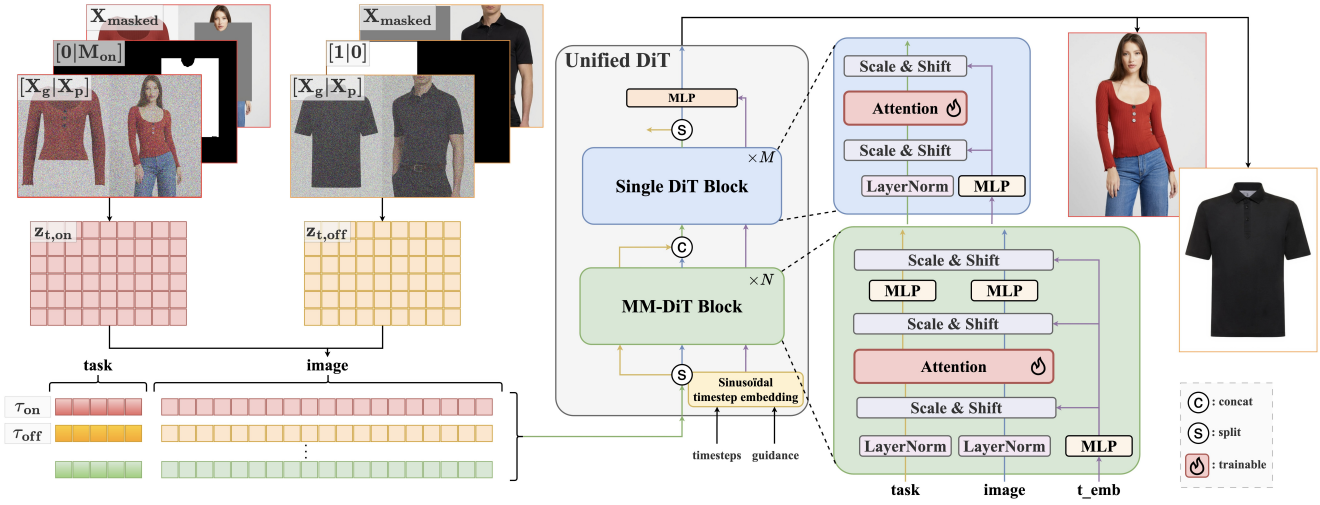

Figure 2. Overview of the proposed model. It is capable of bidirectional virtual try-on/off with a unified transformer.

Figure 3. Attention map visualization. CAT-VTON shows dispersed attention unrelated to the query point, while our model sharply focuses on the corresponding garment (or person) region.

To overcome these limitations, we propose a novel approach that fully exploits the capabilities of diffusion transformers [22]: joint learning of virtual try-on (VTON) and virtual try-off (VTOFF) within a single model. By leveraging the structure of concatenation-based conditioning—where the garment and target images are spatially combined—we design a unified training framework that enables a single diffusion transformer to perform both VTON and VTOFF within the same batch. This is achieved without any additional networks or task-specific losses, while fully preserving the original token layout and model architecture, resulting in a highly efficient and scalable training process.

Unlike prior concatenation-based methods [6] that trivially reconstruct the conditioning garment regions, our joint training framework uses these regions as supervision targets in the VTOFF task, enabling the model to reconstruct the garment from the target image. This dual-role usage reduces redundancy and strengthens garment–target correspondence through bidirectional learning.

Remarkably, our unified diffusion transformer achieves state-of-the-art performance on both VTON and VTOFF benchmarks, outperforming existing models [5, 6, 14, 25, 27, 34] optimized individually for each task. We provide detailed analysis of the enhanced garment–target interaction and validate the effectiveness of our approach through extensive qualitative and quantitative comparisons.

2. Method

2.1. Garment-target correspondence analysis.

In virtual try-on, accurately preserving the shape and details of the input garment on the target person is essential. We analyze the attention maps of transformer to validate garment-target correspondence (Fig. 3). When querying a specific point on the garment or target image, we expect the model to attend to a single, well-aligned location on the corresponding image. However, we observe that the existing method [6] produces dispersed attention across multiple regions, revealing its failure to capture one-to-one spatial alignment between the two domains.

2.2. Try-On & Off via Unified Transformer

Bidirectional try-on/off. Instead of relying on auxiliary components, we propose a fundamentally different approach: training a single transformer to jointly perform both VTON and VTOFF. Our method leverages the structure of concatenation-based conditioning, where both the garment and target images are spatially combined in the input. This formulation naturally enables bidirectional training—from garment to target (VTON) and from target to garment (VTOFF). To implement this, we fix the relative spatial positions of the garment and the person image and vary the inpainting conditioning for each task. As a result, the unified model can generate the garment conditioned on the target, and vice versa, effectively reinforcing garment–target correspondence through joint learning. No-

Figure 4. Qualitative comparison of try-on results with baselines: VITON-HD, DressCode, and in-the-wild (top to bottom).

tably, our method does not introduce any additional visual tokens, external modules or auxiliary loss functions.

Pipeline. Let $\mathbf{X_g} \in \mathbb{R}^{H\times W\times 3}$ denote the standalone garment image and $\mathbf{X_p} \in \mathbb{R}^{H\times W\times 3}$ the target person image. We construct the input by concatenating the two images along the horizontal axis to form the full image, i.e., $\mathbf{X} = [\mathbf{X_g} \mid \mathbf{X_p}] \in \mathbb{R}^{H\times 2W\times 3}$. Task-specific inpainting regions are defined using a binary mask $\mathbf{M} \in \{0,1\}^{H\times 2W}$, which is applied in the image space prior to encoding. For the try-on task, the mask is defined as $\mathbf{M} = [\mathbf{0} \mid \mathbf{M}_{\text{on}}]$, where $\mathbf{M}_{\text{on}}$ masks out the garment region in the person image $\mathbf{X_p}$ while leaving the garment image $\mathbf{X_g}$ unmasked. For the try-off task, the mask is set to $\mathbf{M} = [\mathbf{1} \mid \mathbf{0}]$, masking the entire garment image while keeping the person image unmasked. We then apply the mask to obtain the masked image, $\mathbf{X}_{\text{masked}} = \mathbf{X} \odot (\mathbf{1} - \mathbf{M})$.

Our architecture is based on a latent diffusion model [7, 15], where all denoising operations are performed in the latent space. The full and masked images are encoded into latent representations via a frozen encoder $\mathcal{E}$, yielding $\mathbf{z} = \mathcal{E}(\mathbf{X})$ and $\mathbf{z}_{\text{c}} = \mathcal{E}(\mathbf{X}_{\text{masked}})$. A task token $\tau \in \{\tau_{\text{on}}, \tau_{\text{off}}\}$ is used to distinguish between the try-on and try-off modes and is passed to the transformer as an additional condition.

	VITON-HD [4]				DressCode [20]			
	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓
StableVITON [14]	0.084	0.867	6.85	1.255	0.107	0.905	4.48	1.530
OOTDiffusion [29]	0.096	0.851	6.52	0.896	0.073	0.898	3.95	0.720
IDM-VTON [5]	<u>0.079</u>	**0.881**	6.34	1.322	<u>0.048</u>	**0.923**	3.80	1.201
CatVTON [6]	0.097	0.869	6.14	0.964	0.071	0.901	3.28	0.670
Leffa [34]	0.081	0.872	6.31	1.208	0.060	0.911	3.65	0.709
Ours (w.o. dual)	<u>0.079</u>	0.868	<u>5.80</u>	<u>0.618</u>	0.052	0.910	<u>3.04</u>	<u>0.565</u>
Ours	**0.073**	<u>0.879</u>	**5.52**	**0.406**	**0.045**	<u>0.920</u>	**2.91**	**0.381**

Table 1. Quantitative comparison of try-on task. **Bold** and <u>underline</u> denote the best and second best result, respectively.

Training strategy. We adopt a flow matching formulation [17], where the model learns a time-dependent velocity field that transports samples from the data distribution to noise along a continuous path. Let $\mathbf{z}_0$ be a data latent and $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a sampled noise latent. We define a trajectory $\mathbf{z}_t$ between them and train the denoising model $\epsilon(\cdot)$ to predict the velocity $\frac{d\mathbf{z}_t}{dt}$ at each intermediate point. The unified training objective is given by:

$$\mathcal{L}_{\text{unified}} = \mathbb{E}_{t,\,\mathbf{z}_0,\,\mathbf{z}_1}\left[\left\|\epsilon(\mathbf{z}_t, \mathbf{z}_{\text{c}}, \mathbf{M}, \tau, t) - \frac{d\mathbf{z}_t}{dt}\right\|^2\right]. \quad (1)$$

Figure 5. Qualitative comparison of try-off results with baselines: VITON-HD (top) and in-the-wild (bottom).

	TryOffDiff [25]	TryOffAnyOne [27]	**Ours**
FID↓	28.25	25.20	**10.87**
KID↓	11.42	6.98	**2.57**

Table 2. Quantitative comparison of try-off task.

In our case, we adopt the rectified flow formulation [7, 18], where the trajectory is a straight line between $\mathbf{z}_0$ and $\mathbf{z}_1$, i.e., $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1, \quad \Rightarrow \quad \frac{d\mathbf{z}_t}{dt} = \mathbf{z}_1 - \mathbf{z}_0$. This simplifies training by reducing the target to a constant displacement vector, while aligning with the structure of rectified flows. To enable task-specific adaptation without losing the pre-trained DiT prior, we finetune only the attention modules within each transformer block. This design is well-suited for virtual try-on/off tasks, where accurate spatial reasoning between garment and person is essential.

3. Experiments

3.1. Datasets and experimental setup

We evaluate our method on two standard benchmarks: DressCode [20] and VITON-HD [4], using both qualitative and quantitative metrics. Each dataset contains high-resolution image pairs of in-shop garments and corresponding person images. To assess generalization, we also present qualitative results on in-the-wild images. All outputs are generated at a resolution of 1024×768.

3.2. Qualitative comparison

Figure 4 shows qualitative comparisons between our method and state-of-the-art approaches on the VTON task. VTOFF results are provided in Fig. 5, using the same model for both tasks. Our method generates more coherent and photorealistic results across both VTON and VTOFF.

3.3. Quantitative results

We evaluate visual fidelity and structural consistency using standard metrics. For realism, we report Fréchet Inception Distance (FID) [10] and Kernel Inception Distance (KID) [3]. To assess structural consistency, we use LPIPS [32] and SSIM [26]. As shown in Table 1, 2, our model outperforms existing methods across most metrics.

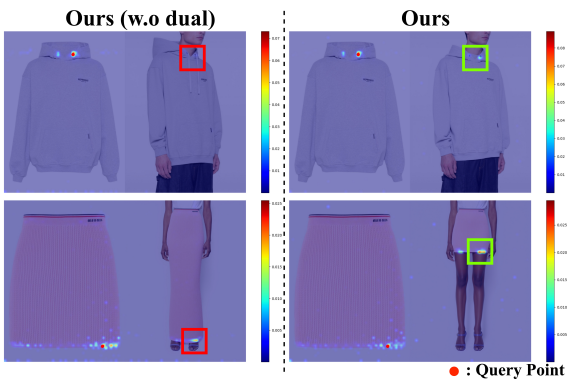

Figure 6. Jointly trained model attends more precisely to relevant regions compared to try-on-only variant.

Training strategy	**# Params**	SSIM↑	LPIPS↓	FID↓	KID↓
Full	11.9B	0.875	0.081	6.35	0.886
Single DiT Blocks	5.38B	0.872	0.078	5.98	0.634
Attention only (ours)	**2.69B**	**0.879**	**0.073**	**5.52**	**0.406**
LoRA	359M	0.843	0.108	6.67	0.906

Table 3. Quantitative comparison of training strategies with varying trainable parameters on VITON-HD [4].

3.4. Ablation study

Effect of dual-task training. To assess the benefit of joint training, we compare our dual-task model with a try-on-only variant. As shown in the last two rows of Table 1, dual-task learning consistently improves performance. As shown in Fig. 6, our model focuses more precisely on the well-aligned corresponding regions, improving spatial alignment.

Effect of trainable parameters. We compare training strategies with different subsets of trainable parameters. As shown in Table 3, our attention-only training achieves the best performance among all methods, including full parameter training, single DiT block training, and low rank adaptation (LoRA) [12]. It effectively captures garment–person interactions while minimizing training overhead.

4. Conclusion

Limitations and future work. While our model is capable of producing photorealistic results by jointly learning try-on and try-off through bidirectional attention, precise control over the garment's fit remains somewhat ambiguous due to the lack of explicit structural or sizing information.

In future work, we plan to incorporate additional cues such as body measurements or garment metadata to improve controllability and personalization.

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 (ECCV)*, 2022. 1

[2] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[3] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021. 4

[4] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3, 4

[5] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3

[6] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, and Xiaodan Liang. CatVTON: Concatenation is all you need for virtual try-on with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025. 1, 2, 3

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024. 1, 3, 4

[8] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *ECCV*, 2021. 1

[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 *ACM International Conference on Multimedia (ACMMM)*, 2023. 1

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 4

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[12] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 4

[13] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv*, 2024. 1

[14] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. StableVITON: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[15] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3

[16] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision (ECCV)*, 2022. 1

[17] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv*, 2023. 3

[18] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv*, 2022. 4

[19] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[20] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4

[21] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *ACM International Conference on Multimedia (ACMMM)*, 2023. 1

[22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 1

[25] Riza Velioglu, Petra Bevandic, Robin Chan, and Barbara Hammer. Tryoffdiff: Virtual-try-off via high-fidelity garment reconstruction using diffusion models. *arXiv*, 2024. 2, 4

[26] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 4

[27] Ioannis Xarchakos and Theodoros Koukopoulos. Tryoffanyone: Tiled cloth generation from a dressed person. *arXiv*, 2025. 2, 4

[28] 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 (NeurIPS)*, 2021. 1

[29] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025. 1, 3

[30] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[31] Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. CAT-DM: Controllable accelerated virtual try-on with diffusion model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[32] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv*, 2018. 4

[33] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[34] Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, Aditya Patel, Tao Xiang, Miaojing Shi, and Sen He. Learning flow fields in attention for controllable person image generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 3

[35] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1