Try-on Diffusion using Diffusion Transformer

Jay Sawant
Halıcıoğlu Data Science Institute
University of California San Diego
La Jolla, CA 92093
jsawant@ucsd.edu

Sunreet Khanna
Jacobs School of Engineering
University of California San Diego
La Jolla, CA 92093
sukhanna@ucsd.edu

Abstract

Virtual try-on aims to render a person wearing a target garment while preserving pose, body shape, and fine-grained appearance details. Recent diffusion-based approaches outperform GAN counterparts yet still rely on UNet backbones that struggle with large geometric misalignments. We revisit the problem through the lens of *Diffusion Transformers* (DiT) and propose architectural and training refinements that faithfully track garment textures while improving geometric fidelity. We leverage the noise-aware parameterization and second-order ODE solver of Karras *et al.* (2022) for the training and sampling process of the diffusion model.

1 Introduction

We study the virtual try-on problem, which seeks to generate a realistic image of a model wearing a target garment given only an initial image of the garment and an image of the model. A related variant uses two input images—one of a source model and one of a different model wearing the garment—to synthesize the source model in that garment. The core challenge is to preserve the model's pose, viewpoint, body shape, and identity while accurately warping the garment to conform to the model and retaining all of its fine visual details. Success in this task offers broad benefits to the fashion industry—enabling immersive online shopping experiences, streamlining garment design, and reducing returns and waste—and also unlocks novel applications in gaming, avatar creation, and virtual reality.

2 Related Work

Image-based virtual try-on methods typically follow a two-stage pipeline: first warping the garment to match the target pose, then blending it onto the person. Early work such as VITON [7] relied on thin-plate-spline (TPS) transformations, which ClothFlow [6] improved by learning dense flow fields. Subsequent models like VITON-HD [3] and HR-VITON [14] focused on generating high-resolution outputs, while SDAFN [1] introduced a single-stage deformable-attention mechanism to enhance alignment. TryOnGAN [15] addresses misalignment through a pose-conditioned StyleGAN2 backbone [12], but this often comes at the cost of attenuating fine garment textures.

More recently, diffusion models [9, 21, 22] have emerged as a stable alternative to GANs [5, 2], offering improved mode coverage and training dynamics. These approaches typically employ a UNet architecture [19] with channel-wise concatenation for conditioning [20]. While effective for aligned tasks such as super-resolution or inpainting, such concatenation struggles to handle the geometric misalignment inherent in try-on scenarios. Moreover, learned warping methods still incur residual misalignments, and latent-space conditioning often fails to preserve detailed garment patterns.

A recent advancement, TryOnDiffusion, decouples garment warping and appearance synthesis into two specialized UNet modules within a diffusion framework [23]. By separating these concerns,

Preprint. Under review.

it achieves both precise geometric alignment and high-fidelity detail preservation—highlighting a promising direction for future virtual try-on systems.

3 Materials and Methods

3.1 Datasets

For our virtual try-on task, we require training pairs in which the same garment is worn by two individuals in different poses, allowing the model to learn pose-invariant garment transfer. To construct such pairs, we utilize the high-resolution VITON-HD [3] dataset and apply the IDM-VITON [4] pipeline to generate clean, pose-varied image pairs while preserving garment fidelity and identity consistency. We also use the official test split from the VITON-HD dataset for evaluation.

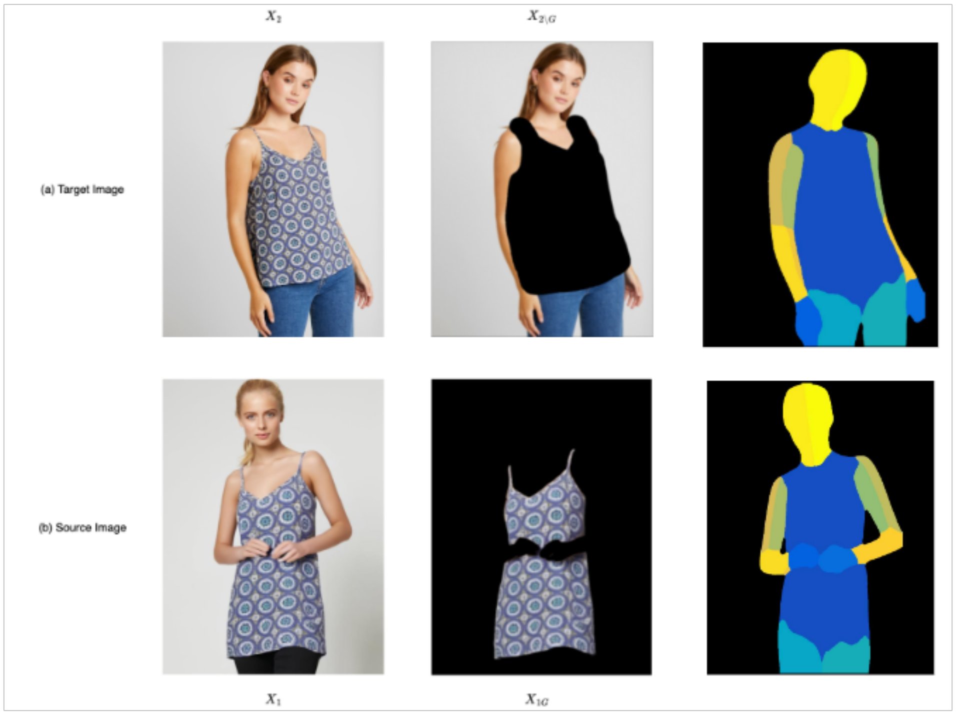

Figure 1: **Pre-processing outputs for a single training pair.** *Top row (a) Target stream.* X_2: original target image; $X_{2\setminus G}$: cloth-agnostic version obtained by removing the garment by using dilated garment mask; S_2: Dense pose extracted from X_2. *Bottom row (b) Source stream.* X_1: source image of another person wearing the same garment; X_{1G}: clean garment mask cropped from X_1; S_1: Dense pose extracted from X_1. These cues form the conditioning tensors $\mathbf{C}_{\text{agnostic}} = [X_{2\setminus G}; S_2]$ and $\mathbf{C}_{\text{garment}} = [X_{1G}; S_1]$ that are fed into the Diffusion Transformer during training.

3.2 Pre-processing

For every training *pair* (X_1, X_2) we run an identical parsing pipeline on each image and then merge their outputs to form the conditioning tensors used in 3.3.

- J_H: 2-D human pose keypoints extracted with DensePose [18].
- S: A single-channel stick-figure rendered from J_H by drawing limbs; we denote the source pose as S_1 and the target pose as S_2.
- X_{1G}: An image of the garment region obtained by garment segmentation method using a pretrained U^2-Net [17], post-processed by closing and largest-component selection to remove holes and spurious fragments.

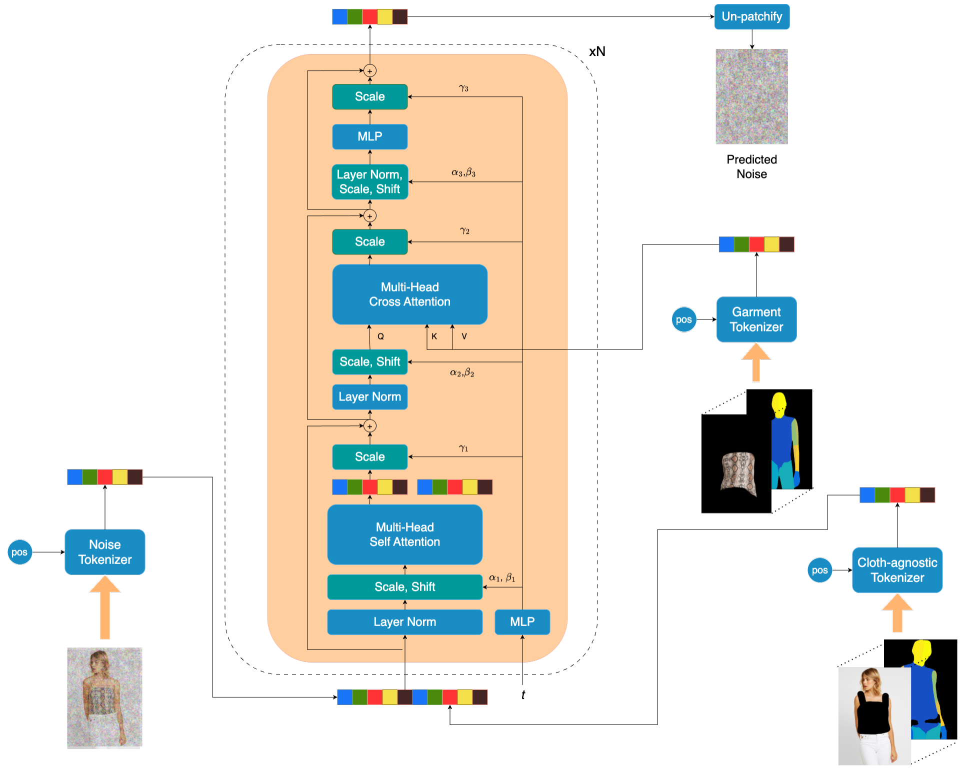

Figure 2: Overview of the proposed geometry-aware DiT. Self-attention operates on noise and cloth-agnostic tokens; a mid-block cross-attention injects garment tokens. FiLM layers (green) modulate tokens with timestep embeddings.

- $X_{2\setminus G}$: A cloth-agnostic human silhouette obtained by dilating the garment mask (radius $\approx$30 px) and subtracting it from a person-parsing map.

We finally compose two multi-channel tensors that serve as input to the transformer:

$$\mathbf{C}_{\text{garment}} = [\, X_{1G} \,;\, S_1 \,], \qquad \mathbf{C}_{\text{agnostic}} = [\, X_{2\setminus G} \,;\, S_2 \,]. \tag{1}$$

Each tensor is split into 16×16 patches, and linearly projected to create the token sequences $\mathbf{z}_{\text{garment}}$ and $\mathbf{z}_{\text{agnostic}}$ consumed in 3.3.

3.3 Model Architecture

We propose a denoising model ϵ_θ is implemented as an N-block Transformer (Figure 2). Each block integrates self-attention over concatenation of noise and cloth-agnostic tokens, cross-attention over the noise and garment tokens, and FiLM-based modulation [16] using the timestep embedding. This design enables the model to conditionally reconstruct the target image from noise with awareness of pose and garment structure.

Training Schema Given a source image X_1 and a target image X_2 of two people wearing the same garment, we denote $X_T = X_2$ and build token sequences as in 3.2. A log-normal noise level σ is drawn and added to $x_0 = X_T$ to obtain $x = x_0 + \sigma\varepsilon$. The network receives $c_{\text{in}}x$ together with $\mathbf{z}_{\text{agnostic}}$ (self-attention) and $\mathbf{z}_{\text{garment}}$ (cross-attention) and is optimized to predict the EDM [11] target (3.4).

3.4 Noise-aware Forward and Reverse Processes

We closely follow the *EDM* formulation of Karras [11]. We set $\sigma_{\text{data}} = 0.66$ for images in the range $[-1, 1]$. For every training sample we draw a *log-normal* noise level

$$\sigma \sim \exp\big(\mathcal{N}(\mu, \sigma^2)\big), \quad \mu = P_{\text{mean}} = -1.2, \; \sigma = P_{\text{std}} = 1.2. \tag{2}$$

A noised sample is $x = x_0 + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$.

Re-weighting. Signal-to-noise factors are

$$c_{\text{in}} = (\sigma^2 + \sigma_{\text{data}}^2)^{-1/2}, \qquad c_{\text{skip}} = \sigma_{\text{data}}^2/(\sigma^2 + \sigma_{\text{data}}^2), \qquad c_{\text{out}} = \sigma\, c_{\text{in}}\, \sigma_{\text{data}}. \tag{3}$$

The network receives $c_{\text{in}}x$ and is trained to predict

$$y_{\text{target}} = \frac{x_0 - c_{\text{skip}}x}{c_{\text{out}}}. \tag{4}$$

Weighted MSE loss. Following EDM we minimise a *scale-aware* mean-squared error

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0,\varepsilon,\sigma}\Big[w(\sigma)\,\|\,\epsilon_\theta(c_{\text{in}}x, \sigma, c) - y_{\text{target}}\|_2^2\Big], \tag{5}$$

with

$$w(\sigma) = \frac{\sigma^2 + \sigma_{\text{data}}^2}{\sigma^2\,\sigma_{\text{data}}^2}. \tag{6}$$

This weighting prevents small-σ samples from dominating the gradient signal.

Second-order ODE solver. During inference we adopt the ρ-controlled noise schedule

$$\sigma(t) = \big[\sigma_{\text{max}}^{1/\rho} + t\,(\sigma_{\text{min}}^{1/\rho} - \sigma_{\text{max}}^{1/\rho})\big]^{\rho}, \quad t \in [0, 1], \tag{7}$$

with $\rho = 7$, $\sigma_{min} = 0.01$ and $\sigma_{max} = 80$. Given consecutive noise levels $\sigma_i > \sigma_{i+1}$ the solver updates

$$x_{i+1} \;=\; x_i + (\sigma_{i+1} - \sigma_i)\,\frac{x_i - \hat{\varepsilon}_\theta(x_i, \sigma_i, c)}{\sigma_i} - \tfrac{1}{2}\,(\sigma_{i+1} - \sigma_i)^2\,\partial_\sigma\Big[\tfrac{x_i - \hat{\varepsilon}_\theta}{\sigma}\Big], \tag{8}$$

where $\hat{\varepsilon}_\theta$ employs classifier-free guidance with weight w (Eq. 9). Empirically, $K{=}256$ noise levels suffice for 512 × 384 images.

Classifier-Free Guidance (CFG). We employ classifier-free guidance [10] to improve generation fidelity by interpolating between conditional and unconditional noise predictions:

$$\hat{\epsilon}_\theta(x_t, c) = (1 + w)\cdot\epsilon_\theta(x_t, c) - w\cdot\epsilon_\theta(x_t, \varnothing), \tag{9}$$

where w is the guidance weight and $\epsilon_\theta(x_t, \varnothing)$ is the model prediction with dropped conditionals.

4 Implementation Details

All experiments are run with the code available at https://github.com/jay6101/p2p_tryon

Input resolution and tokenization. Images are resized to 512×384. With a patch size of 8×8 each view yields $64{\times}48 = 3\,072$ spatial tokens. At every denoising step the self-attention stream sees 3 072 noise tokens plus 3 072 cloth-agnostic tokens (6 144 total), while the cross-attention stream receives the 3 072 garment tokens.

Diffusion Transformer. The denoiser consists of 16 DiT blocks (hidden width 512, MLP width 2 048, 8 attention heads).

Diffusion setup. EDM schedule (Eq. 7) with $\sigma_{\min} = 0.01$, $\sigma_{\max} = 80$, $\rho = 7$. Conditioning is dropped with probability 0.1 and classifier-free guidance uses weight $w = 3.0$. Sampling employs the second-order solver with 128 noise levels.

Optimizer and schedule. AdamW with learning rate 3×10^{-4}, warm-up for 2 000 steps, and cosine decay thereafter. EMA decay is 0.999.

Batching and training length. Physical batch size 2; training runs for $600K$ iterations with FID computed at regular checkpoints.

All the training was implemented using 1 GPU with 24 GB VRAM (RTX 4090).

5 Results and analysis

Evaluation Metrics We use the Fréchet Inception Distance (FID) [8] to quantify the realism of generated try-on images. FID computes the distance between multivariate Gaussians fitted to Inception features of real and generated images. Let (μ_r, Σ_r) and (μ_g, Σ_g) be the means and covariances of real and generated features, respectively. Then:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\right). \tag{10}$$

We compute the FID of our synthesized images against a test set of 1983 images. We synthesize 1983 garment transfer images of size 512x512 and compare them against the real test set images to obtain an FID of **27.7574**.

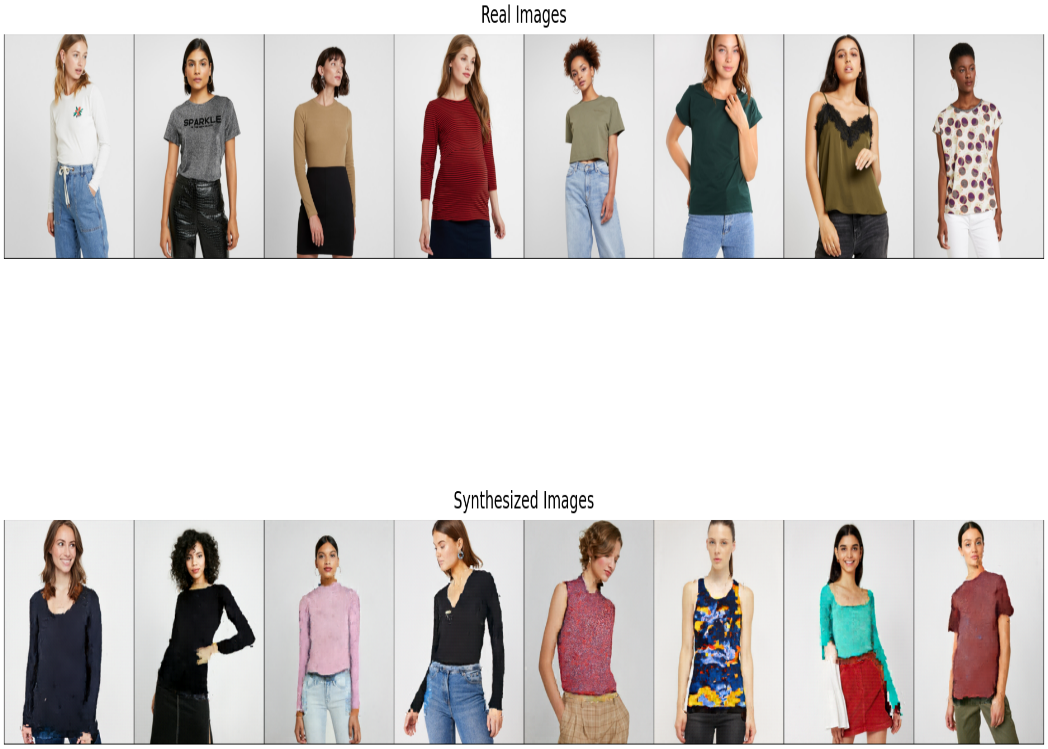

Figure 3: Sample images in comparitive analysis fir FID computation.

5.1 Qualitative Results and Observations

We present initial qualitative outputs of our virtual try-on framework (4, 5) showcasing synthesized results on the training set using early-stage diffusion-transformer models. These models are still under active development, and due to limited GPU resources, we have not yet been able to train high-capacity variants with smaller patch sizes or for extended epochs. Consequently, while some high-frequency garment textures and intricate details are not fully preserved, the current results already display several promising aspects.

Pose Capture and Occlusion Handling The model exhibits strong performance in capturing the target body pose, including the generation of challenging regions such as occluded necks, hair, and arms. These aspects often require nuanced attention across multiple input modalities, and the current model manages to maintain spatial coherence in many such instances.

Garment Structure and Color Reproduction Across most samples, the synthesized garments maintain good structural integrity and accurate reproduction of color properties. The garment is often well warped and aligned with the target pose, preserving silhouette and fabric fall. This suggests that the model has learned reasonable spatial correspondence between the source and target.

Design Pattern Fidelity The framework shows an emerging ability to reconstruct basic garment design patterns. In particular, abstract or colorful patterns are generally well represented, likely because of their distributed texture and broader coverage across training samples. However, reconstruction quality degrades when patterns involve fine-grained, semantically specific content such as text, brand logos, or localized motifs.

Loss of Fine Detail Some high-frequency features—such as frills, embroidery, buttons, or sharp textures—are often missed or blurred in the output. This can be attributed to the relatively coarse patch size used in the transformer encoder, as well as the limited training duration, which may not suffice to capture fine visual cues.

Impact of Computational Constraints Due to hardware limitations, model training was restricted to roughly 100K iterations per configuration—substantially below the intended training schedule of up to 100 million iterations. This severely restricts the model's ability to fully converge and adapt to the complexity of garment detail and pose diversity. We strongly believe that with extended training—especially under longer schedules and finer granularity in visual input—the model's capacity to capture detailed texture and structural fidelity will substantially improve.

Limited Hyperparameter Exploration Our current results are derived from a narrow configuration set due to constrained experimentation bandwidth. We have only partially explored variations in patch size, number of transformer layers, attention head counts, cross-attention design, and hierarchical scale-specific attention mechanisms. These architectural components are known to significantly influence performance in vision-transformer-based pipelines, and we anticipate that a broader sweep would yield notable gains in fine detail reconstruction, texture continuity, and generalization.

6 Future Work

In future iterations, we plan to improve the quality of cloth-agnostic representations by generating more semantically meaningful and spatially accurate maps, which are essential for preserving pose and identity while enabling realistic garment transfer. Additionally, we aim to integrate a fine-tuning framework built on top of a pretrained diffusion backbone, such as Flux.1-fill [13], to leverage strong generative priors and accelerate convergence, ultimately improving texture fidelity and structural consistency with fewer resources.

References

[1] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows, 2022.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019.

[3] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization, 2021.

[4] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild, 2024.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[6] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R. Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[7] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network, 2018.

[8] 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*, volume 30, 2017.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

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

[11] Tero Karras, Miika Aittala, Samuli Laine, Timo Herva, and Jaakko Lehtinen. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

[14] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions, 2022.

[15] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation, 2021.

[16] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017.

[17] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, October 2020.

[18] Iasonas Kokkinos R{iza Alp Güler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. 2018.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[20] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2022.

[21] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[22] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[23] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets, 2023.

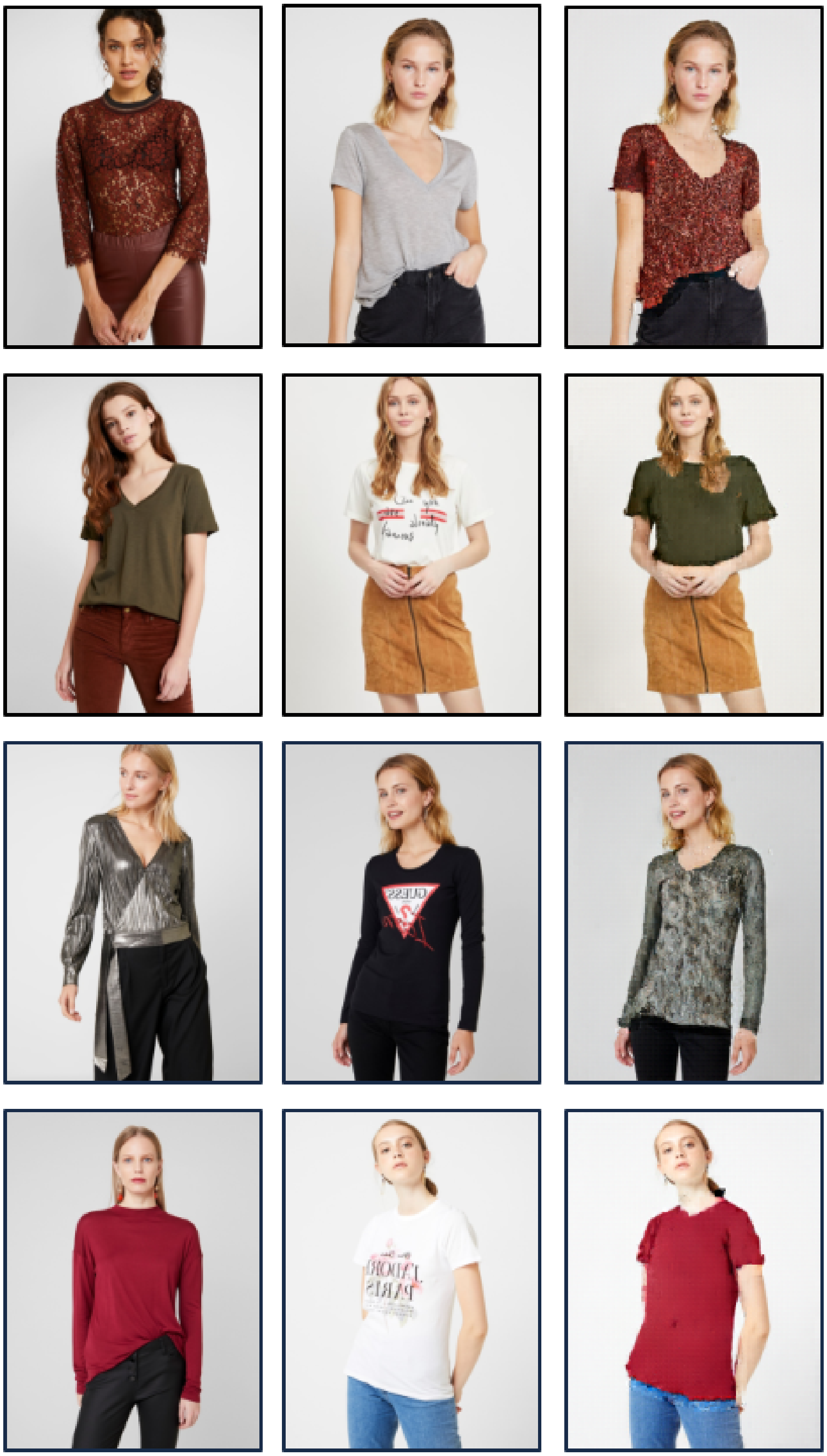

Figure 4: From left to right, Column I shows the model image, Column 2 shows the target garment to be transferred, and Column 3 shows the generated image

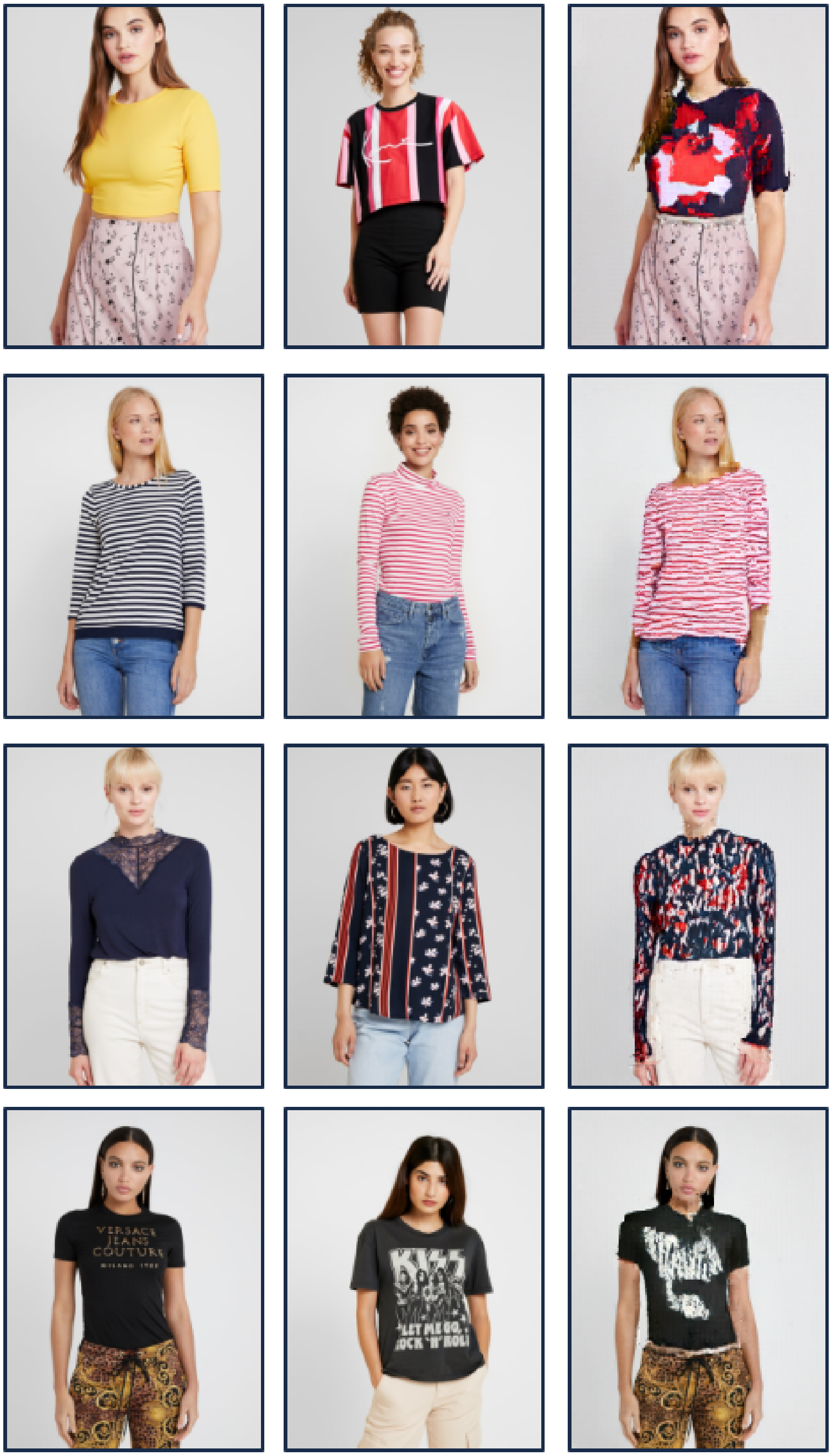

Figure 5: From left to right, Column I shows the model image, Column 2 shows the target garment to be transferred, and Column 3 shows the generated image