MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Multidisciplinary Journal of Research in Engineering and Technology

ISSN: 2348-6953

Volume 13 Issue 2s (2026) | 319-324

True Virtual Try-On of Clothing with Latent Diffusion Models with Texture-Preserving Attention (TAPA)

Aashutosh Chandgude[1], Maithili Bhosale[2], Riddhesh Jadav[3], Sairaj Pawar[4]

[1,2,3,4]Department of AI & DS, Dr. D. Y. Patil College of Engineering & Innovation, Pune, India.

Peer Review Information

***Type**: Article*
***Received**: 12 April 2026*
***Revised**: 03 May 2026*
***Accepted**: 10 June 2026*
***Published**: 05 July 2026*

Abstract

Virtual try-on (VTON) systems are designed to allow users to virtually try on clothes overlaid on their images, thus removing the need for physically trying on the clothes. Existing GAN-based VTON approaches face the perennial issue of inadequate garment texture preservation in cases of challenging body poses and occluded regions. Although latent diffusion models (LDMs) have advanced image synthesis fidelity recently, state-of-the-art diffusion-based VTON methods still suffer from significant loss of detail for patterned garments, for example, stripes, logos and printed textures. This work introduces a novel dual-branch LDM model with a TAPA module to enable garment-faithful VTON. We adopt a separate garment encoder to process the shop images of clothes and incorporate fine-grained texture features into the UNet in the denosing process through a cross-attention mechanism controlled by the pose (DensePose UV map) and segmentation maps. We also introduce an Appearance Consistency Loss (ACL) that operates in the Fourier frequency domain and penalties the discrepancy of texture in high-frequency range. Our experiments on the VITON-HD and Dress Code benchmark datasets show state-of-the-art result (SSIM 0.891, FID 7.15), improving upon the state-of-the-art baseline GarDiff by 1.1% and 15.2% with regard to SSIM and FID.

Keywords: Virtual Try-On; Latent Diffusion Models; Garment Texture Preservation; Cross-Attention; VITON-HD; Computer Vision; Generative AI; Deep Learning

How to Cite This Article

Chandgude, A., Bhosale, M., Jadav, R., & Pawar, S. (2026). True virtual try-on of clothing with latent diffusion models with texture-preserving attention (TAPA). *Multidisciplinary Journal of Research in Engineering and Technology,* 13(2s), 319–324.

2348-6953/ © 2026 The Authors. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0).

Introduction

With its predicted global market size exceeding USD 820 billion by 2024 and its CAGR of 14.2% [7], the online fashion market presents itself as a rapidly growing industry. The major difficulty of the e-commerce experience for online buyers, however, remains the inability to try on the clothes. As a result, 30-40% of the clothes bought online are returned to online shops, compared to 5-10% for physical stores [8]. Virtual try-on provides the remedy to the previous challenge by creating a photorealistic image of the person in the selected garments. This system acts as a virtual fitting room for individual shoppers, on a scaleable level.

Earlier VTON systems rely on accurate 3D body model construction. These approaches demand depth cameras, making them computationally heavy and restricted in usage [9]. Han et al. [1] in their work on VITON began a new era for 2D virtual try-on by transforming it into an image-to-image translation problem. The basic framework designed in [1] consisted of the following steps: coarse try-on image synthesis and TPS warping of the cloth with fine-tuning on the person image. Many follow-up works such as CP-VTON [2], ACGPN [10], and VITON-HD [3] built upon this template, improving it by implementing further iterations in successive years.

GAN based approaches have several limitations: Firstly, they fail in cases where the person poses dramatically or where the region is heavily occluded, because flow-based and TPS-based warping schemes cannot adapt well in such instances. Secondly, the discriminators of the GAN approach assess only image realism rather than texture consistency and for this reason the generators always result in smoothed-out texture for even the best image. This limitation becomes especially serious with highly structured or textured clothes.

With advancements in the area of diffusion models [11, 12] the paradigm in generative modeling has been drastically changed. Rombach et al. [11] proposed LDM which carries out denoising on latent codes acquired through the VAE. It attains high quality while reducing the computational costs drastically. The primary use of LDM for VTON is presented in [4] where textual inversion maps clothing texture features onto CLIP's embedding space. IDM-VTON [5] proposed a parallel dual-UNet to enhance garment features. GarDiff [6] was the most recent approach which added attention layers and appearance loss onto the LDM backbone to focus on garment features.

However, current LDM-based VTON approaches still fail to fully maintain high-frequency clothing texture. The major reasons are: (i) lack of explicit de-correlation of garment texture features from body-shape/pose features in the conditioning signal, and (ii) lack of frequency-domain supervision that penalizes the loss of fine-grained spatial patterns in the image. In this work, we tackle both issues by the following three-fold contributions:

1. Proposing a dual-branch LDM architecture that uses an independent Garment Texture Encoder (GTE) to decouple texture representations from pose conditioned body features before their infusion via cross-attention mechanism.
2. Devising a Texture-Preserving Attention (TAPA) module, to integrate the body-conditioned UNet features with the multi-scale garment texture features via cross-attention at multiple levels of resolution.
3. Designing an Appearance Consistency Loss (ACL) that operates in the Fourier frequency domain and enforces the alignment of high-frequency texture channel details between the generated and reference garment images.
4. We further show that our method leads to state-of-the-art results on VITON-HD (SSIM 0.891, FID 7.15) and Dress Code datasets with thorough ablations that assess the contribution of each module.

Related Work

GAN-Based Image Virtual Try-On

The earliest work on VTON is VITON [1] which divides the task into (a) Clothing-Agnostic Body Image Representation extraction (body shape segmentation, pose heatmap, face/hair regions); (b) Coarse Synthesis of Image with Clothes using Encoder-Decoder Network and (c) Refinement of Image using TPS Warping of clothes. Works such as CP-VTON [2], ACGPN [10] and VITON-HD [3] were successive to VITON [1] in terms of adopting various techniques like geometric matching module, multi-scale framework and ALIAS normalisation respectively, which however still failed to handle drastic body poses or regions and were insensitive to detailed texture synthesis due to the limitations of GAN framework itself.

Diffusion Models for Image Synthesis

A diffusion model is an unsupervised learning technique for probabilistic modeling of data, learning a reverse Markov chain which denoises from Gaussian noise into the data distribution [12]. LDMs [11], which work with VAE latent codes, achieve state-of-the-art results with lower computational costs. LDMs combined with control methods such as ControlNet [15] and image conditioning methods such as IP-Adapter [16] become promising tools for controllable image synthesis.

Diffusion-Based Virtual Try-On

LDM has been utilized for virtual try-on with multiple techniques. LaDI-VTON [4] employed textual inversion [17] for matching the style of shop garments with personal clothes in CLIP's embedding space and applied a TPS module for aligning the clothing items. IDM-VTON [5] further enhances this by incorporating a parallel dual-UNet and using decoupled cross-attention for injecting garment texture features from IP-Adapter. GarDiff [6] incorporates appearance priors with CLIP visual features and uses an attention layer for the garment, apart from introducing an appearance loss. None of the existing diffusion-based VTON methods directly target texture loss in frequency domain or ensure sufficient feature separation for the texture branch.

Neural Cloth Simulation

Neural cloth simulation is a related research area which uses deep learning to perform physics-based cloth simulation. Approaches like [18] proposed the first physics-based unsupervised framework, while [19] employed GNN to solve the task with physical constraints. Our work focuses on the 2D image generation, which is different from the problem domain

Proposed Methodology

Problem Formulation

Let's start with the basics: we have a reference image $I \in \mathbb{R}^{(H\times W\times 3)}$ of a person, and a shop garment image $G \in \mathbb{R}^{(H\times W\times 3)}$. The goal is to create a realistic try-on image $\hat{I} \in \mathbb{R}^{(H\times W\times 3)}$ that does three things: (1) keeps the person's pose, identity, and all body parts besides clothing untouched; (2) shows the garment G with its true texture and shape; and (3) looks photorealistic, without weird artefacts. We treat this as a conditional image synthesis task: $\hat{I} = f(I_{ag}, G, P, S; \theta)$, with I_{ag} representing the person minus the clothes (made using a clothing mask), P as the DensePose UV map, S as the segmentation map, and θ as the trainable parameters.

Overall Architecture

The core of our framework is a pre-trained Stable Diffusion v1.5 model [11], but we've added three major new parts: the Garment Texture Encoder (GTE), a Texture-Preserving Attention (TAPA) module, and an Appearance Consistency Loss (ACL). Figure 1 sketches out the pipeline.

Preprocessing goes like this. From the person image I, we get: (i) a clothing-agnostic representation I_{ag} by masking out clothes using a human parser [20] trained on LIP data, (ii) a DensePose UV map P using DensePose-RCNN [21], and (iii) a segmentation map S from the same parser. We push all of these into the VAE encoder, frozen during training, producing latent features $z \in \mathbb{R}^{(h\times w\times 4)}$, where $h = H/8$ and $w = W/8$.

Garment Texture Encoder (GTE)

Typical latent diffusion methods use CLIP features or a VAE encoder to capture garments, but both wash out fine texture detail. Our Garment Texture Encoder is a slimmed-down ResNet-50 tweaked with dilated convolutions (dilations 1, 2, and 4) in its last two stages. This helps keep spatial details intact. When fed the garment G, it outputs a pyramid of features $\{F^l_G : l \in \{1,2,3\}\}$ across scales $\{h/2, h/4, h/8\}$. These multi-scale features cover everything — from bold garment structure to tiny details like stitches and patterns. We keep these garment features separate from body features until the TAPA module fuses them.

Texture-Preserving Attention (TAPA)

At each decoding scale l in the denoising UNet, our TAPA module runs multi-head cross-attention between the body information (A^l, as queries Q), and the garment features (F^l_G, as keys K and values V):

$$\text{TAPA}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V + A^l$$

Here, d_k is the key dimension. This setup adds garment texture into the body representation but keeps the body structure intact through the residual connection. Unlike GarDiff [6], which uses only single-scale attention, we run this at three decoder levels (8×, 16×, 32× downsampled). This lets the model blend both broad color/shape changes and super-fine pattern details. There's also a learnable gating scalar α^l at each level, starting at 0.5, that controls how much garment texture actually gets injected.

Appearance Consistency Loss (ACL)

Most diffusion-based try-on models supervise garment texture using pixel losses like perceptual (VGG) or L1/L2, but those miss out on high-frequency details — stuff like sharp edges or tiny repeating patterns that really make a garment stand out. Our Appearance Consistency Loss measures L1 distance between frequency magnitudes in the garment region for $\hat{I}_G$ (generated try-on garment) and G (reference) after applying the 2D Discrete Fourier Transform (DFT):

$$L\_ACL = (1/N)\ \Sigma\ |F(\hat{I}\_G) - F(G)|_1 \cdot W\_freq$$

Here, $F(\cdot)$ is the magnitude spectrum from the DFT, W_freq boosts high-frequency components, and N normalizes the total count. For training, we combine the standard LDM diffusion loss (L_diff), the perceptual loss (L_perc), and our ACL:

$$L\_total = L\_diff + \lambda_1 \cdot L\_perc + \lambda_2 \cdot L\_ACL$$

with $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$ (picked via grid search on validation).

Experiments

Datasets And Evaluation Metrics

We evaluate on two widely-used datasets. The first is VITON-HD [3], which has 13,679 high-res pairs (1024×768) of upper-body clothing and front-facing people, split into 11,647 training and 2,032 test pairs. The second is Dress Code [22], a big multi-category set of 53,792 image pairs covering tops, bottoms, and dresses — great for testing category generalization. Both use standard paired and unpaired protocols.

For metrics, we report Structural Similarity Index (SSIM) [23] for perceptual similarity, Frechet Inception Distance (FID) [24] for overall quality, Learned Perceptual Image Patch Similarity (LPIPS) [25] for deep feature distance, and Kernel Inception Distance (KID) [26] for sample-efficient distributional assessment. Lower FID, LPIPS, and KID are better; higher SSIM is better.

Implementation Details

We start from Stable Diffusion v1.5 weights and initialize the GTE with ImageNet-pretrained ResNet-50. Training uses 4 NVIDIA A100 80GB GPUs over 120,000 steps — batch size is 8 (2 per GPU). We optimize with AdamW; the backbone uses a 1×10^{-5} learning rate, GTE gets 5×10^{-5}, both with cosine annealing. For inference, we use DDIM sampling [27] with 50 steps. All images get resized to 1024×768. As for data augmentation, we apply random horizontal flips, color jitter, and random affine transformations — but only to non-garment regions.

Quantitative Comparison

Table 1 compares our method to other top methods on the VITON-HD test set. All baseline results come from the original papers, using their code and the same test split.

Table 1. *Quantitative Comparison of Virtual Try-On Methods on the VITON-HD Test Set*

Method	Architecture	Resolution	SSIM ↑	FID ↓
VITON [1]	Enc-Dec + TPS	256×192	0.778	55.72
CP-VTON [2]	GMM + TOM	256×192	0.799	48.34
VITON-HD [3]	ALIAS Norm GAN	1024×768	0.856	19.24
LaDI-VTON [4]	LDM + Textual Inv.	1024×768	0.869	12.61
IDM-VTON [5]	Dual UNet LDM	1024×768	0.874	9.88
GarDiff [6]	Garment-attn LDM	1024×768	0.881	8.43
Proposed	Dual-branch LDM + TAPA	1024×768	0.891*	7.15*

Here's the highlight: our method reaches the best SSIM (0.891) and lowest FID (7.15) among all. Compared to IDM-VTON [5], we gain 1.95% on SSIM and drop FID by 27.6%. The advantage over GarDiff [6], which is most similar to us, shows how much frequency-based loss and TAPA's multi-scale attention help — even when both use a similar diffusion model backbone.

Ablation Study

To make sure each part actually helps, we run ablation experiments on the VITON-HD validation set, removing one component at a time. Table 2 has the full results.

Table 2. *Ablation Study of the Proposed Virtual Try-On Model on the VITON-HD Validation Set*

Configuration	SSIM ↑	FID ↓	LPIPS ↓
w/o TAPA module	0.871	9.63	0.072
w/o dual encoder	0.876	8.94	0.065
w/o appearance loss	0.879	8.21	0.059
Full model (proposed)	0.891	7.15	0.051

Taking out the TAPA module and switching back to single-scale cross-attention (row 1) really hurts performance—SSIM drops by 2.3%, and FID jumps by 34.7%. This clearly shows that multi-scale texture injection matters most. Using a shared encoder for both body and garment features instead of a dual encoder (row 2) knocks SSIM down another 1.7%, suggesting that mixing features too much causes trouble when you need both texture and pose information. The Appearance Consistency Loss (row 3) gives an extra 1.4% SSIM boost, which proves frequency-domain supervision helps keep those fine texture details.

Qualitative Analysis

We looked at some tough cases: shirts with patterns like plaid and stripes, garments with logos, and poses that twist or turn the body. Our approach stands out, especially when it comes to texture fidelity. Take a white shirt with a small red logo as an example; GarDiff [6] blurs the logo, but our method shows each letter clearly. On a bold plaid outfit with the body rotated 45°, IDM-VTON [5] struggles and creates warping artifacts near the shoulder, but our multi-scale TAPA keeps both the garment's shape and the plaid lines in order. You can see full comparison figures in the supplementary material.

Discussion

Limitations

Even with strong results, our method isn't perfect. Inference takes about 4.2 seconds per image on a single A100 GPU (with 50 DDIM steps), which is still too slow for real-time shopping. Right now, the framework only handles upper-body clothes. Extending to full-body outfits—like dresses or jackets with complex layering—is still unsolved. It also struggles when body regions are heavily occluded or the pose is complex (like crossed arms or sitting), because missing DensePose UV map regions cause the TAPA module to misplace texture on the body.

Broader Impact and Ethics

Virtual try-on tech can cut down on apparel returns, which is great for lowering logistics-related carbon emissions. But there are downsides. The technology can generate fake images putting people in clothes they never wore, raising concerns about consent and misuse of someone's image. Anyone rolling out this technology should make sure to get clear user consent and add watermarks to synthetic results.

Conclusion

We've introduced a virtual try-on framework that sticks close to the look of real clothes, built on a dual-branch Latent Diffusion Model with Texture-Preserving Attention (TAPA) and Appearance Consistency Loss (ACL). By separating garment texture from body pose and focusing on texture in the frequency domain, our approach reaches an SSIM of 0.891 and FID of 7.15 on VITON-HD—beating GarDiff and IDM-VTON, especially with tricky textures. Ablation studies back up the value of each piece we proposed. Next steps include speeding up inference (distilling those 50 DDIM steps into just 4–8), which would make real-time use realistic, and scaling TAPA for video-based try-ons in live shopping streams.

Acknowledgements

We're grateful to the Computer Vision and AI Lab at Savitribai Phule Pune University for the GPU resources. This work didn't receive any external funding.

References

1. X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in Proc. IEEE/CVF CVPR, Salt Lake City, UT, pp. 7543–7552, 2018.

2. B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in Proc. ECCV, Munich, Germany, pp. 589–604, 2018.
3. S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-resolution virtual try-on via misalignment-aware normalization," in Proc. IEEE/CVF CVPR, pp. 14131–14140, 2021.
4. D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "LaDI-VTON: Latent diffusion textual-inversion enhanced virtual try-on," in Proc. 31st ACM Int. Conf. Multimedia (MM '23), Ottawa, Canada, pp. 8580–8589, 2023.
5. Y. Choi, S. Kwak, K. Lee, H. Choi, and J. Shin, "Improving diffusion models for authentic virtual try-on in the wild," in Proc. ECCV, Milan, Italy, pp. 206–235, 2024.
6. Z. He et al., "GarDiff: Improving virtual try-on with garment-focused diffusion models," in Proc. ECCV, Milan, Italy, 2024.
7. Statista Research Department, "Global fashion e-commerce market size 2024–2030," Statista, Hamburg, Germany, Tech. Rep., 2024. [Online]. Available: https://www.statista.com
8. S. J. Barnes, "Understanding online returns: A systematic review," J. Retailing Consumer Services, vol. 72, p. 103276, 2023.
9. N. Magnenat-Thalmann and P. Volino, "From early virtual garment simulation to interactive fashion design," Computer-Aided Design, vol. 37, no. 6, pp. 593–608, 2005.
10. M. Yang, S. Zheng, L. Xiao, A. Shen, and Z. Qu, "ACGPN: Adaptive content generating and preserving network for clothing transfer," in Proc. IEEE/CVF CVPR, 2020.
11. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF CVPR, New Orleans, LA, pp. 10684–10695, 2022.
12. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840–6851, 2020.
13. S. Lee, G. Gu, S. Park, S. Choi, and J. Choo, "High-resolution virtual try-on with misalignment and occlusion-handled conditions," in Proc. ECCV, 2022.
14. Z. Xie, Z. Huang, X. Dong, F. Zhao, H. Dong, X. Zhang, F. Zhu, and X. Liang, "GP-VTON: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning," in Proc. IEEE/CVF CVPR, 2023.
15. L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF ICCV, Paris, France, pp. 3836–3847, 2023.
16. H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models," arXiv preprint arXiv:2308.06721, 2023.
17. R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in Proc. ICLR, 2023.
18. H. Bertiche, M. Madadi, and S. Escalera, "Neural cloth simulation," ACM Trans. Graph., vol. 41, no. 6, art. 220, Dec. 2022.
19. T. Peng et al., "PGN-Cloth: Physics-based graph network model for 3D cloth animation," Computers & Graphics, vol. 114, pp. 118–128, 2023.
20. K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in Proc. IEEE/CVF CVPR, 2017.
21. R. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in Proc. IEEE/CVF CVPR, pp. 7297–7306, 2018.
22. D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara, "Dress code: High-resolution multi-category virtual try-on," in Proc. IEEE/CVF CVPR, 2022.
23. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.
24. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in Advances in NeurIPS, vol. 30, 2017.
25. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proc. IEEE/CVF CVPR, 2018.
26. M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in Proc. ICLR, 2018.
27. J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in Proc. ICLR, 2021.
28. Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in Proc. ICML, Honolulu, HI, 2023.