International Journal of Electrical and Electronics Research ISSN 2348-6988 (online)
Vol. 14, Issue 2, pp: (1-9), Month: April - June 2026, Available at: www.researchpublish.com

AI-Powered Virtual Outfit Try-On System for Sustainable Fashion and Sizing Accuracy

Nishchay Sinha[1], Pinank Trivedi[2], Yashraj Verma[3], Deepak S. Shete[4]

[1,2,3] Department of Electronics and Telecommunication Engineering

Thakur College of Engineering and Technology, Mumbai, India

[1,2,3] Authors, [4]Guide

DOI: https://doi.org/10.5281/zenodo.19483948

Published Date: 09-April-2026

Abstract: **This paper presents DM-VTON, a diffusion-model-based virtual try-on (VTON) system designed to generate high-quality, photo-realistic visualizations of clothing on human subjects. The system simulates how garments fit and drape on the human body under real-world conditions, accounting for diverse poses, lighting variations, and body shapes. It integrates advanced artificial intelligence techniques including diffusion models, IP-Adapters, DensePose, and semantic segmentation to achieve accurate and realistic garment rendering. The model was trained on benchmark datasets VITON-HD and DressCode, enabling robust handling of both paired and unpaired try-on scenarios. Key contributions include dense pose mapping, agnostic image generation, and a Gradio-based user interface supporting high-resolution inference at 1024×768 pixels. The modular and scalable architecture effectively tackles persistent challenges such as occlusion handling, body shape variation, and lighting inconsistency. This work advances the fields of virtual try-on, fashion synthesis, and AI-driven apparel visualization, with significant applicability to e-commerce, fashion design, and sustainable retail.**

Keywords: **Virtual Try-On, Diffusion Models, DensePose, IP-Adapters, Semantic Segmentation, Fashion Synthesis, E-Commerce, Sustainable Fashion, Agnostic Image Generation, Garment Rendering.**

I. INTRODUCTION

The global fashion and e-commerce industries are witnessing a rapid surge in demand for realistic and accessible virtual try-on (VTON) systems. These systems aim to provide consumers with an immersive virtual experience of wearing clothing prior to purchase, thereby boosting consumer confidence and significantly reducing costly product return rates. According to recent industry reports, online fashion return rates can reach as high as 40%, imposing severe environmental and financial burdens on retailers and consumers alike. A robust and photorealistic VTON system has the potential to drastically mitigate this issue by enabling users to accurately visualize how a garment will appear on their unique body.

Traditional VTON solutions, which often rely on basic 2D image overlays or early generative models such as Generative Adversarial Networks (GANs), frequently fail in real-world scenarios. Their limitations include poor occlusion handling, inability to accurately simulate fabric deformation across diverse body poses, and inadequate adaptability to a wide variety of body shapes and sizes. These shortcomings have limited their adoption and restricted consumer trust in virtual fitting technologies.

To address these shortcomings, this paper introduces DM-VTON (Diffusion Model for Virtual Try-On), a state-of-the-art framework leveraging advanced diffusion models. The system is augmented with IP-Adapters for intelligent garment feature management, DensePose for detailed body mapping, and human parsing for improved structural consistency. The result is a photorealistic, flexible, and scalable virtual try-on system suitable for deployment across e-commerce platforms, fashion design workflows, and sustainable retail applications.

A. Motivation and Applications

The motivation for this project stems from the evident shortcomings of existing VTON technologies. Despite functional utility, current systems lack the realism and robustness required for mainstream consumer adoption. The DM-VTON system addresses these gaps by delivering a highly convincing virtual try-on experience. In online retail, the system allows customers to visualize how garments look on their specific body type, leading to increased conversions and reduced returns. For fashion designers, it offers rapid digital prototyping, enabling swift design iterations without requiring physical samples. Crucially, DM-VTON contributes to sustainable consumer behaviour by minimizing the environmental burden of physical product returns and overproduction.

Beyond e-commerce, potential applications include fashion education, virtual styling assistants, AR/VR fashion experiences, and accessible clothing design for people with disabilities. These diverse use cases affirm the broad societal and commercial relevance of the proposed system.

B. Problem Definition

The primary problem addressed in this work is the lack of photorealism and robustness in current VTON systems when confronted with uncurated, real-world conditions. Existing systems often fail to accurately simulate garment behaviour including fabric drape, wrinkle formation, lighting adaptation, and interaction with complex body shapes and poses. A particularly significant challenge is occlusion handling, such as when a person's arms are folded across their body, requiring the system to intelligently reconstruct missing garment regions. Additionally, the ideal system must seamlessly compose a person's image with a standalone clothing image while preserving user identity, including facial features and body shape.

Current approaches also struggle to generalize across diverse ethnic body types and non-standard poses encountered in real-world photography. Addressing these challenges requires a fundamentally different generative paradigm beyond GANs.

C. Objectives and Scope

The central objectives of this project are:

- Develop a state-of-the-art VTON system using diffusion models for highly realistic garment visualization.
- Integrate DensePose mapping and semantic segmentation for accurate body and garment representation.
- Support both paired and unpaired try-on scenarios with high visual fidelity at 1024×768 pixel resolution.
- Train and validate on large-scale benchmark datasets (VITON-HD and DressCode).
- Deploy via a user-friendly, real-time Gradio-based interactive interface.
- Demonstrate quantitative and qualitative superiority over prior VTON approaches.

D. Paper Organisation

The remainder of this paper is organised as follows. Section II presents a comprehensive literature review of existing VTON methods. Section III describes the proposed system and its methodology. Section IV covers system design and requirements. Section V details the implementation and experimental setup. Section VI presents results and discussion. Section VII draws conclusions and outlines future directions.

II. LITERATURE REVIEW

The domain of virtual try-on has advanced considerably through the convergence of computer vision, deep learning, and generative modelling. Early systems employed basic 2D image overlays with significant limitations in realism and body adaptability. Arulananth et al. [1] introduced a Python-based smart trial room using OpenCV and a webcam; while cost-effective, it suffered from inaccurate garment fit due to 2D overlay constraints. Shadrach et al. [2] developed a more interactive VTON model with improved garment alignment and real-time webcam support, though adaptability to varied body shapes remained limited.

Kim et al. [3] proposed a style-controlled clothing synthesis model using convolutional neural networks (CNNs) to disentangle geometric structure from clothing style, enabling manipulation without distorting the wearer's posture. Mohammadi and Kalhoor [4] provided a comprehensive review of AI-based fashion synthesis, highlighting contributions of GANs, pose estimation, and semantic segmentation. Sani et al. [5] introduced a style synthesis framework focused on

texture preservation and pose-adaptive garment rendering. Akmal [6] identified key advancements in clothing detection and VTON platforms, emphasising scalability and dataset diversity.

De Almeida [7] found that perceived realism, ease of use, and trustworthiness are critical factors for consumer adoption of VTON systems. Marelli et al. [8] explored web-based VTON using OpenPose and semantic segmentation, focusing on performance and browser compatibility. Jadhav et al. [9] proposed AI-driven virtual model generation for automated fashion catalogue creation. A systematic review by Chen et al. [10] revealed that while photorealism has significantly improved, challenges remain in tactile feedback simulation and mobile optimisation. These prior works collectively define the motivation and scope for the DM-VTON system proposed herein.

Diffusion models represent a newer generation of deep generative models, originally introduced for image synthesis tasks [Ho et al., 2020]. Compared to GANs, diffusion models offer more stable training dynamics, higher output diversity, and superior photorealism, making them ideally suited for the complex image synthesis required in VTON. Recent works such as Paint-by-Example and InstructPix2Pix have demonstrated the versatility of diffusion architectures in image editing, while LDM (Latent Diffusion Models) have made high-resolution synthesis computationally tractable. DM-VTON builds upon these foundations by conditioning the diffusion process on body pose and garment features.

A. Limitations and Gap Analysis

Existing VTON systems exhibit several critical limitations that motivate the present work:

- Systems based on 2D overlays lack photorealism and fail to simulate accurate garment fit, resulting in unnatural-looking outputs.
- GAN-based approaches, while more advanced, often produce visible artifacts, mode collapse, and lack training stability, limiting their generalisation to diverse inputs.
- More sophisticated models still struggle with occlusion handling, support for diverse body types, and visual coherence in dynamic environments.
- Challenges also persist in mobile platform optimisation, AR/VR integration, and the diversity of available training datasets.
- Most existing models do not adequately address full-body garment try-on, focusing instead on upper-body clothing only.

These gaps form the motivation for the DM-VTON framework, which is designed specifically to address them through a principled combination of diffusion-based generation, IP-Adapter-guided feature conditioning, and dense body mapping. Table I summarises the comparative positioning of prior works.

TABLE I: COMPARATIVE LITERATURE REVIEW

Ref.	Method	Backbone	Key Limitation
[1]	Smart Trial Room	OpenCV 2D Overlay	Inaccurate fit, no 3D body modelling
[2]	Webcam VTON	2D Overlay + CNN	Limited body shape adaptability
[3]	Style-controlled CNN	CNN disentanglement	Restricted to upper-body garments
[4]	AI Fashion Review	GAN + Pose	Artifacts in complex poses
[5]	Style Synthesis	GAN + Texture	Texture loss in occluded regions
[8]	Web-based VTON	OpenPose + Segmentation	Browser performance bottleneck
Ours	DM-VTON	Diffusion + IP-Adapter	Higher inference time vs. 2D methods

III. PROPOSED SYSTEM AND METHODOLOGY

The DM-VTON system is built upon a structured and advanced pipeline designed to realistically simulate how a garment appears on a person in real-world environments. The pipeline begins with two primary inputs: a person image and a clothing item image. The system generates an "agnostic image" of the user—where the original clothing is masked while identity and pose are retained—and simultaneously applies DensePose mapping to create a detailed body map. These representations, along with extracted garment features, are fed into a diffusion model that iteratively denoises a noisy image conditioned on body structure and clothing information, producing a high-resolution, photo-realistic try-on output.

A. Input Acquisition and Preprocessing

The system accepts two inputs: a static person image or live webcam feed, and a clothing item image. During preprocessing, semantic segmentation is applied using a state-of-the-art human parsing network to create an agnostic image—the person's representation without their original garment, with identity, hair, face, and body shape preserved. This step is critical for ensuring that the diffusion model generates the new garment in the correct body region without conflating the original clothing with the target.

Concurrently, DensePose mapping generates a dense 3D mesh-like representation of the user's body surface, capturing body contours, surface orientation, and spatial geometry critical for realistic garment projection. DensePose partitions the body surface into 24 semantically meaningful regions, each encoded with UV coordinates that allow precise garment warping and alignment.

Image normalisation, resizing to 1024×768, and colour space conversion are also performed during preprocessing to ensure consistent input quality across diverse real-world photographs. Noise and illumination artefacts in input images are attenuated through adaptive histogram equalisation.

B. Garment Feature Extraction

Visual and optionally textual (caption-based) features are extracted from the clothing image. A pre-trained CLIP visual encoder processes the garment image to produce a rich, semantically aware feature vector that encodes key garment attributes including colour, fabric texture, structural patterns, and design details. These features are essential for accurate garment rendering on the target person.

Optionally, a garment captioning module generates a natural language description of the clothing item, which is encoded using CLIP's text encoder and fused with the visual features. This multi-modal conditioning enables more nuanced garment representation, particularly for items with complex patterns or structural details such as buttons, collars, or embroidery.

C. Diffusion Model and IP-Adapter Integration

The core of DM-VTON is a diffusion generative model based on the latent diffusion model (LDM) architecture. Unlike GANs, diffusion models iteratively denoise a random noise image conditioned on structured inputs, producing highly consistent and artefact-free outputs. The denoising process is guided by a UNet backbone operating in the latent space of a pre-trained VAE, enabling high-resolution synthesis at reduced computational cost.

The model is conditioned on three complementary inputs: the agnostic image (person without original garment), the DensePose body map (structural body geometry), and extracted garment features (clothing appearance and texture). IP-Adapters are integrated into the denoising process at multiple UNet attention layers to intelligently blend garment features, ensuring that details such as fabric texture, colour patterns, and garment structure are faithfully preserved in the final synthesis. Cross-attention mechanisms within the IP-Adapter allow the model to selectively attend to different garment regions during generation.

D. Agnostic Image Generation

A critical preprocessing step is the generation of the agnostic image—a version of the input person image in which the clothing region is masked and inpainted with a neutral skin-tone background. This process involves: (1) semantic segmentation to identify the garment region at the pixel level; (2) body-part-aware masking that selectively removes the clothing while preserving exposed skin, face, hair, and accessories; and (3) inpainting using a dedicated network to produce a plausible body surface beneath the removed garment. The agnostic image provides the diffusion model with a clean body canvas onto which the target garment is synthesised.

E. Algorithm

The core DM-VTON algorithm proceeds as follows:

1. Input: Acquire the person image P and target clothing image C.
2. Preprocessing: - Generate an agnostic image A from P by removing clothing regions using human parsing. Apply DensePose to P to obtain body representation B. Resize and normalize P and C to a fixed resolution (e.g., 1024×768).
3. Feature Extraction: Extract visual features F from C using CLIP encoder; optionally generate caption and encode as text features F_t.

4. Encoding: Encode A into latent space z_A via VAE encoder; initialise noisy latent z_T ~ N(0, I).
5. Denoising: Feed z_T, z_A, B, F, F_t to the LDM UNet; iteratively denoise z_T for T steps, conditioning on B and F at each step via cross-attention.
6. IP-Adapter Integration: Blend F into UNet attention layers during each denoising step to preserve garment texture and pattern fidelity.
7. Decoding: Decode final denoised latent z_0 via VAE decoder to produce output image O at 1024×768.
8. Rendering: Display the generated try-on image O using a user interface (e.g., Gradio), with options for visualization and export.

IV. SYSTEM DESIGN AND REQUIREMENTS

A. Functional Requirements

The DM-VTON system must fulfil the following functional requirements to meet its intended use cases:

- Accept a person image (static or live webcam) and a clothing item image as input.
- Accurately segment and mask original clothing while preserving identity and pose (agnostic image generation).
- Generate a detailed DensePose map capturing body contours and orientation.
- Extract visual and optional text-based features from the garment image.
- Synthesise a new image of the person realistically wearing the target garment using a diffusion model conditioned on body and garment information.
- Handle both paired and unpaired person-garment combinations.
- Generate high-resolution output images at 1024×768 pixels.
- Provide an interactive Gradio-based UI for real-time try-on, pose adjustment, and image export.

B. Non-Functional Requirements

- Performance: Generate try-on images in near-real-time (≤ 5 seconds per inference on recommended hardware) for practical usability.
- Scalability: Handle large numbers of concurrent users and clothing items without performance degradation through modular microservice architecture.
- Reliability: Produce consistent, high-quality results across diverse lighting conditions, body types, and garment categories.
- Usability: Intuitive Gradio interface accessible to non-technical users with no installation required beyond a web browser.
- Security: Ensure user data privacy and image security in production deployments through encrypted communication and ephemeral storage.
- Maintainability: Modular architecture enabling independent update and replacement of system components.

C. Technology Stack

TABLE II: TECHNOLOGY STACK

Component	Technology / Role
Generative Model	Diffusion Model – iterative denoising for high-fidelity image synthesis
Body Mapping	DensePose – dense human body surface estimation capturing 3D geometry and pose
Garment Masking	Semantic Segmentation – pixel-level classification to isolate original clothing

Feature Blending	IP-Adapters – transfer garment attributes (colour, texture, pattern) into generation
Training Data	VITON-HD and DressCode – large-scale benchmark datasets
User Interface	Gradio – interactive web-based UI for real-time try-on and image export
Frameworks	PyTorch / TensorFlow; OpenCV for computer vision operations

D. System Architecture

The DM-VTON system follows a client-server architecture. The client is the Gradio-based web interface running in a standard browser, and the server hosts the core AI pipeline. The server comprises several interconnected modules: the Input Handler (manages uploads and webcam streams), the Preprocessing Module (semantic segmentation and DensePose), the Feature Extraction Module (CLIP-based visual and text encoding), the Core Diffusion Model with IP-Adapters (LDM UNet denoising), and the Output Generator (VAE decoding and image rendering). The modular design ensures each component can be independently developed, tested, and replaced, supporting long-term maintainability and extensibility.

Communication between modules uses an internal REST API with JSON payloads for metadata and binary streams for image data. GPU memory management is handled through dynamic batching and cache clearing between inference calls. Cloud deployment leverages containerisation (Docker) for reproducibility and horizontal scaling.

V. IMPLEMENTATION AND EXPERIMENTAL SETUP

A. Datasets

The DM-VTON model was trained and validated using two established benchmark datasets. A summary is provided in Table III.

TABLE III: DATASET SUMMARY

Dataset	Images	Category	Resolution	Paired
VITON-HD	13,679	Upper-body	1024×768	Yes
DressCode	53,792	Upper/Lower/Full	1024×768	Yes/No

VITON-HD is a high-resolution dataset comprising 13,679 paired person-garment image sets designed for upper-body try-on tasks, with images at 1024×768 pixel resolution. Each sample includes a front-facing person photograph, a flat-lay clothing image, an agnostic person image, a DensePose map, and a human parsing map, providing a comprehensive set of paired training signals.

DressCode is a diverse multi-category dataset encompassing 53,792 paired samples across upper-body, lower-body, and full-body garment categories. This dataset introduces significant variety in garment types (shirts, trousers, dresses), body poses, and demographic diversity. Together, these datasets provide the necessary variety in clothing styles, body shapes, and poses to ensure model robustness and generalisability across real-world conditions.

B. Hardware and Software Setup

The training and inference of the diffusion model require substantial computational resources. The software stack includes PyTorch 2.0 as the primary deep learning framework, with Hugging Face Diffusers library used for LDM implementation. OpenCV 4.8 is employed for computer vision preprocessing operations, and Gradio 4.x is used for user interface development.

On the hardware side, training and high-resolution inference require a powerful GPU with sufficient VRAM. The recommended configuration is an NVIDIA A100 80GB GPU alongside a modern 16-core CPU and at least 64 GB of system RAM. For inference only, an NVIDIA RTX 3090 (24 GB VRAM) or equivalent is sufficient. Cloud-based GPU services (Google Cloud TPU and AWS EC2 P4) were utilised during training to manage computational overhead cost-effectively.

Training was performed for 200,000 steps with a batch size of 4, using the AdamW optimiser with a learning rate of 1×10^{-4} and cosine annealing schedule. Mixed-precision training (fp16) was used to reduce memory consumption by approximately 40%.

C. Training Procedure

The DM-VTON training pipeline follows a two-stage approach. In the first stage, the VAE and CLIP encoders are frozen, and only the UNet denoising backbone is fine-tuned on the VITON-HD dataset using a combination of reconstruction loss (L2 in latent space) and perceptual loss (LPIPS). This stage establishes the base garment-conditioned denoising capability.

In the second stage, the IP-Adapter layers are introduced and trained jointly with the UNet on the combined VITON-HD and DressCode datasets. The IP-Adapter parameters are optimised using a separate learning rate (5×10^{-5}) to avoid overfitting while preserving the pre-trained LDM priors. Data augmentation including random horizontal flipping, brightness jitter, and cropping is applied to improve generalisation.

D. Feasibility Analysis

A three-dimensional feasibility study was conducted prior to implementation. Technically, the project is feasible as all core technologies are well-established with open-source implementations available in the research community. The use of pre-trained LDM checkpoints (Stable Diffusion) reduces training time and resource requirements substantially. Economically, open-source frameworks and pre-trained models significantly reduce development costs, with a clear ROI potential through e-commerce integration, where even a 10% reduction in return rates represents substantial savings for retailers. Operationally, the Gradio-based interface ensures accessibility for non-technical users, requiring no specialised training.

VI. RESULTS AND DISCUSSION

A. Outputs and Outcomes

The primary outcome of this project is the successful development and deployment of the DM-VTON system, capable of generating photo-realistic outfit visualisations at 1024×768 pixel resolution. In the generated outputs, the target clothing item appears naturally draped and shaded in accordance with the person's body shape and orientation, while the user's identity remains fully intact. The system successfully supports both paired and unpaired try-on scenarios, demonstrating high flexibility and realism.

Key functional features—including dense pose mapping, agnostic image generation, and garment captioning—collectively contribute to the superior quality of generated images. The Gradio interface enables users to upload person and garment images, trigger inference, view the result side-by-side with the original, and export the output in PNG or JPEG format.

B. Quantitative Performance Evaluation

The system is evaluated using established image quality metrics: Fréchet Inception Distance (FID, lower is better), Structural Similarity Index Measure (SSIM, higher is better), and Learned Perceptual Image Patch Similarity (LPIPS, lower is better). Table IV presents a quantitative comparison of DM-VTON against representative GAN-based and 2D overlay methods.

TABLE IV: QUANTITATIVE PERFORMANCE COMPARISON

Metric	DM-VTON	GAN-Based	2D Overlay
Photorealism (FID ↓)	12.4	28.7	67.3
SSIM Score (↑)	0.87	0.72	0.45
LPIPS (↓)	0.09	0.21	0.53
Occlusion Handling	High	Moderate	Low
Body Shape Support	Diverse	Limited	Very Limited
Output Resolution	1024×768	512×512	256×256
Inference Time (sec)	~4.2	~1.8	~0.3

DM-VTON achieves an FID of 12.4, significantly outperforming GAN-based methods (28.7) and 2D overlay approaches (67.3), confirming its superior photorealistic output quality. The SSIM score of 0.87 and LPIPS of 0.09 further corroborate the high structural fidelity and perceptual quality of the generated images. The trade-off is a slightly higher inference time (~4.2 seconds) compared to simpler methods, which is acceptable for non-real-time e-commerce applications but is targeted for optimisation in future work.

C. Qualitative Evaluation

A qualitative user study was conducted with 30 participants who rated DM-VTON outputs on a 5-point Likert scale across three dimensions: photorealism, garment fit accuracy, and identity preservation. DM-VTON achieved mean scores of 4.3/5 for photorealism, 4.1/5 for garment fit accuracy, and 4.5/5 for identity preservation, confirming strong user acceptance. Participants particularly noted the realistic fabric draping and the absence of the artefacts commonly observed in GAN-based outputs.

Visual inspection of outputs across diverse body types, skin tones, and garment categories confirms robust generalisation. The system handles complex garment structures such as patterned fabrics, ruffled textures, and structural elements (e.g., collars, buttons) with high fidelity. Occlusion scenarios—such as crossed arms—are handled gracefully through the agnostic image inpainting pipeline.

D. Discussion

The results demonstrate a significant advancement over traditional VTON approaches. Unlike 2D overlay-based methods and earlier GAN architectures—which struggle with occlusions, pose variations, and clothing deformation—the DM-VTON system effectively addresses these challenges through its modular, scalable architecture. The integration of IP-Adapters within the diffusion denoising process proves particularly effective for preserving fine-grained garment details such as fabric texture and design patterns.

The system's ability to maintain high photorealism across diverse body types and complex poses confirms its potential for real-world deployment in fashion and e-commerce. Limitations identified include higher inference latency compared to GAN-based methods, occasional hallucination of fine print text on garments, and reduced performance on highly reflective or metallic fabrics. These limitations define clear directions for future improvement.

From a sustainability perspective, the DM-VTON system has the potential to significantly reduce fashion industry waste. By enabling consumers to virtually try on clothing before purchase, the system can reduce impulse buying and lower the return rate, directly decreasing the carbon footprint associated with return logistics and overproduction.

VII. CONCLUSION

This paper has presented DM-VTON, a sophisticated virtual try-on system representing a significant advancement in the field of AI-driven fashion synthesis. The system implements a comprehensive AI pipeline integrating diffusion models, IP-Adapters, DensePose mapping, and semantic segmentation to produce photo-realistic garment visualisations at 1024×768 resolution. By training on large-scale benchmark datasets VITON-HD and DressCode, the system achieves robust performance across paired and unpaired try-on scenarios, demonstrating strong generalisation to diverse body types, poses, and garment categories.

Quantitative evaluation confirms that DM-VTON achieves state-of-the-art photorealism with an FID of 12.4, SSIM of 0.87, and LPIPS of 0.09, substantially outperforming prior GAN-based and 2D overlay methods. Qualitative user studies further validate strong user acceptance across dimensions of realism, fit accuracy, and identity preservation. The development of a user-friendly Gradio interface democratises access to this advanced technology, making it suitable for deployment in consumer-facing e-commerce applications without requiring specialised hardware or technical expertise.

The modular architecture successfully resolves persistent challenges in occlusion handling, body shape diversity, and lighting inconsistency, confirming DM-VTON's readiness for integration into e-commerce, fashion design, and sustainable retail platforms. The system's contribution to sustainable fashion—through reduced return rates and more informed purchasing decisions—represents a meaningful intersection of AI research and environmental responsibility.

Future work will explore several promising directions: (1) AR/VR integration for immersive try-on experiences; (2) mobile platform optimisation through model distillation and quantisation; (3) tactile and fit simulation to complement visual output; (4) multi-garment try-on support (full outfits including accessories); (5) real-time inference optimisation targeting sub-second latency; and (6) user feedback loop mechanisms to enable personalised style recommendations. These directions collectively position DM-VTON as a foundation for next-generation AI-powered fashion platforms.

ACKNOWLEDGEMENT

The authors sincerely thank their project guide Mr. Deepak S. Shete, Assistant Professor, for his invaluable guidance, mentorship, and unwavering support throughout this work. Heartfelt gratitude is extended to the Principal Dr. B.K. Mishra

and HOD Dr. Payel Saha of the Department of Electronics and Telecommunication at Thakur College of Engineering and Technology, Mumbai, for their constant encouragement and institutional support. The authors also acknowledge the contributions of industry experts who provided feedback during development, and their respective families for their enduring motivation and support.

REFERENCES

[1] T. S. Arulananth, B. Kumar, K. Dasari, S. V. S. Prasad, C. Mahitha, and V. Manohar, "Python based smart trial room," International Research Journal of Engineering and Technology, vol. 9, no. 12, Dec. 2022.

[2] F. D. Shadrach, M. Santhosh, S. Vignesh, S. Sneha, and T. Sivakumar, "Smart virtual trial room for apparel industry," International Journal of Computer Applications, vol. 184, no. 11, Apr. 2022.

[3] B.-K. Kim, G. Kim, and S.-Y. Lee, "Style-controlled synthesis of clothing segments for fashion image manipulation," IEEE Transactions on Multimedia, vol. 21, no. 7, pp. 1681–1695, Jul. 2019.

[4] S. O. Mohammadi and A. Kalhoor, "Smart fashion: A review of AI applications in virtual try-on & fashion synthesis," International Journal of Computer Applications, Nov. 2021.

[5] S. R. Sani, S. M. R. Mallireddy, N. K. R. Renati, and P. L. S. Surya, "Style synthesis: AI-powered dress try-on experience," IEEE Access, May 2024.

[6] A. Akmal, "A comprehensive survey on AI-driven fashion technologies: Clothing detection, recommendation systems, and virtual try-on solutions," ACM Computing Surveys, Nov. 2024.

[7] M. I. G. De Almeida, "Consumers' acceptance of artificial intelligence virtual try-on systems when shopping apparel online," Journal of Retailing and Consumer Services, vol. 58, 2021.

[8] D. Marelli, S. Bianco, and G. Ciocca, "Designing an AI-based virtual try-on web application," in Proc. IEEE Conf. Computer Vision Workshops, May 2022.

[9] A. V. Jadhav, M. A. Bhosle, V. U. Shinde, A. P. Hend, and A. C. Karve, "AI-driven virtual model generation for fashion catalog creation," IEEE Trans. Emerging Topics in Computing, May 2025.

[10] C. Chen, J. Ni, and P. Zhang, "Virtual try-on systems in fashion consumption: A systematic review," Journal of Retailing and Consumer Services, vol. 78, Dec. 2024.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF CVPR, pp. 10684–10695, 2022.

[13] A. Radford, J. W. Kim, C. Hallacy et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021.

[14] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in Proc. IEEE/CVF CVPR, pp. 7297–7306, 2018.

[15] Y. Li, C. Huang, and C. C. Loy, "Dense intrinsic appearance flow for human pose transfer," in Proc. IEEE/CVF CVPR, 2019.