IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025

DOI: 10.17148/IJARCCE.2025.14246

ENHANCED CAT-DM: OPTIMIZED DIFFUSION-BASED VIRTUAL TRY-ON

Dr.O.Aruna[1], Muvvala Hemanth[2], Arimanda Ma Dushyanth Reddy[3] , Ramisetty Lokesh[4] , Pamulapati Sivaiah[5]

Professor, CSE (Cybersecurity, IOT Including Blockchain Technology),

Vasireddy Venkatadri Institute of Technology, Guntur, India[1]

Student, CSE (Cybersecurity, IOT Including Blockchain Technology),

Vasireddy Venkatadri Institute of Technology, Guntur, India[2]

Student, CSE (Cybersecurity, IOT Including Blockchain Technology),

Vasireddy Venkatadri Institute of Technology, Guntur, India[3]

Student, CSE (Cybersecurity, IOT Including Blockchain Technology),

Vasireddy Venkatadri Institute of Technology, Guntur, India[4]

Student, CSE (Cybersecurity, IOT Including Blockchain Technology),

Vasireddy Venkatadri Institute of Technology, Guntur, India[5]

Abstract: Enhanced CAT-DM is an advanced virtual try-on system that combines diffusion models with GAN-based initialization to attain greater realism, efficiency, and controllability. It builds on the Garment-Conditioned Diffusion Model (GC-DM) and incorporates DINO-V2, a top-performing self-supervised vision model, for fine-grained, pixel-level garment representations. ControlNet is also used to improve conditioning accuracy such that garments fit naturally onto body shapes. In order to speed up the normally time-consuming sampling process of diffusion models, we propose a truncation-based acceleration method that leverages a GAN-synthesized coarse image as an initial guess. This largely minimizes the number of sampling steps needed without compromising high-fidelity garment details. In addition, Poisson blending is employed to blend the synthesized garments into the target person's image with seamless transitions and realistic texture conservation. Extensive assessments on benchmark datasets show that Enhanced CAT-DM beats current virtual try-on techniques in terms of higher LPIPS, SSIM, and CLIP-I scores, which confirm its superiority in retaining high-level details, structural features, and garment semantics. All these innovations render Enhanced CAT-DM highly appropriate for real-time, high-fidelity virtual try-on scenarios, filling the gap between AI-based garment synthesis and real-world usability in fashion and e-commerce sectors.

Keywords: Virtual Try-On, Diffusion Models, Generative Adversarial Networks (GANs), Garment-Conditioned Diffusion Model (GC-DM), ControlNet, DINO-V2, Truncation-Based Acceleration, Poisson Blending.

I. INTRODUCTION

Image-based Virtual Try-On [2],[6],[10] is an important computer vision research field that deals with conditional person image generation. The main goal is to generate realistic images of people wearing target garments while maintaining key features like identity, pose, and garment-related features like texture, pattern, and fit. The task is quite challenging owing to the necessity of smooth garment integration and fine-grained texture preservation. Classic methods tend to use Generative Adversarial Networks (GANs) [20], usually deforming the in-shop clothes images to conform to a subject's body first before generating the final result. These methods suffer from complicated poses, occlusions, and failing to preserve textures, tending to produce bent clothes and artificial layouts in generated images.

Recently, diffusion models have gained attention for their exceptional generative capabilities and their ability to produce high-quality, high-fidelity images. Unlike GANs [20], diffusion models generate detailed and realistic outputs by progressively refining an image from noise. Despite their advantages, the application of diffusion models in virtual try-on systems faces two major obstacles: controllability and computational cost. These models involve many sampling steps, which result in heavy computational requirements that are not suitable for real-time use. Furthermore, it is difficult to obtain accurate garment shape and texture alignment because diffusion models tend to focus more on global coherence than on fine textures.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025
DOI: 10.17148/IJARCCE.2025.14246

To overcome these issues, the Enhanced CAT-DM (Clothing-Aware Transformer with Diffusion Models) [19] proposes several innovations upon the Garment-Conditioned Diffusion Model (GC-DM) [19]. One of the key improvements is the incorporation of DINO-V2 [1] for conditioning and feature extraction and ControlNet-based conditioning, which supports pixel-accurate reconstruction of garments. DINO-V2 [1] improves texture preservation, and ControlNet offers accurate garment placement via spatial control signals, leading to more natural and aesthetically pleasing try-on results. Additionally, CAT-DM [19] uses a truncation-based acceleration approach to minimize inference time dramatically. Rather than starting with raw Gaussian noise, the model uses a coarse image generated using a GAN [20] as a point of departure for the diffusion process. This composite method reduces the number of needed diffusion steps with minimal loss in output fidelity, resulting in improved speed and reduced computational complexity for image generation.

Aside from its novel diffusion process, CAT-DM [19] also uses Optimized Poisson Blending to make the image more realistic by removing perceptible artifacts and providing seamless garment transitions. In contrast to basic image stitching techniques, Poisson Blending maintains natural boundaries and shading, leading to artifact-free and professionally completed virtual try-on images. Exhaustive comparisons with perceptual and semantic consistency metrics, such as LPIPS [8], SSIM [4], and CLIP-I [17], show that CAT-DM [19] outperforms current techniques in visual realism and computational efficiency. Due to its evenly weighted strategy to high fidelity, velocity, and responsiveness, CAT-DM [19] promises significant potential in uses for e-commerce, personal fashion advice, and useful real-world virtual try-on systems.

II. LITERATURE SURVEY

Morelli et al. and Cucchiara.R [14], investigates a new virtual try-on system, "Dress Code," that can visualize garments at high resolution in various categories. The method uses sophisticated computer vision methods to allow realistic garment fitting on digital avatars. Using deep neural networks, the system naturally incorporates clothing pieces into user-specified images without losing fine-grained details like textures and patterns. The model enhances the realism and accuracy of virtual try-ons by a large extent, providing a leap in online shopping experience through the resolution of issues such as occlusions, pose variations, and garment flexibility.

Xie et al. [16], introduces GP-VTON, an end-to-end virtual try-on framework capable of supporting diverse clothing styles and poses. Through the integration of local-flow modelling and global-parsing learning, the framework supports better garment alignment and natural visualization. GP-VTON is superior at maintaining garment details and transforming them into intricate human poses, solving issues such as varied clothing deformations and occlusions. This collaborative learning method breaks the limits of virtual try-on systems, opening doors to general-purpose use in e-commerce and digital fashion.

Zhu et al.[18], in their research, unveils Tryon Diffusion, a revolutionary virtual try-on system using a dual U-Net architecture to provide high-fidelity garment visualization. The system integrates two domain-specific U-Nets: one for accurate clothing alignment and the other for creating realistic image details. Through the incorporation of diffusion models, Tryon Diffusion is superior at maintaining garment textures and fitting clothing to various human poses with minimal distortion. This method raises the bar for photorealistic virtual try-ons, closing the gap between accuracy and visual realism in online fashion use cases.

Chen et al. [21], provides a size-sensitive virtual try-on system that prioritizes correct garment fit through the Clothing-Oriented Transformation Try-On Network (COTTON). This technique specifically caters to size challenges by involving garment and human body measurements in the transformation process. Through concentration on clothing size and orientation, the system ensures proper garment alignment and fit over diverse body shapes. The model outperforms visually coherent try-on generation while retaining delicate garment features, and is an important leap forward in size-aware virtual try-on systems towards tailored fashion experience.

Li et al. [13], present a new virtual try-on system employing pose-garment key point guided inpainting to realize realistic garment visualization. This method fits clothes onto humans by detecting garment and body key landmarks, promoting accurate fit and pose consistency. The inpainting method is highly effective at filling gaps, covering occlusions, and maintaining garment texture and detail. Utilizing pose and garment key points, the approach maximizes visual coherence and flexibility of virtual try-ons, providing a strong solution for online fashion applications with dynamic pose variations.

Yang et al.'s [11] paper, shows a Full-Range Virtual Try-On system utilizing a Recurrent Tri-Level Transform (RTLT) to support smooth virtual try-ons with varying clothing styles and human poses. Tri-level transform combines garment alignment, texture smoothing, and pose adjustment to support high-quality outcomes for intricate clothing variations. The recurrent component continually enhances garment fit and perceptual realism, accommodating difficulties such as occlusions and significant pose drifts. This holistic solution improves the scalability and realism of virtual try-on systems, which makes it applicable to different online fashion contexts.

Bai et al. [5], parades a Single Stage Virtual Try- On model with Deformable Attention Flows that streamlines and improves the process of virtual try-on.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025
DOI: 10.17148/IJARCCE.2025.14246

This method obviates the requirement for multi-stage pipelines by merging garment alignment and texture synthesis into a single architecture. Dynamic deformable attention flows capture spatial relationships between humans and clothes, providing accurate fit and realistic visualization. Streamlining the process, the framework results in high efficiency and quality, solving problems such as pose variation and intricate garment deformations for virtual try-on systems.

Minar et al.'s study [15], demos CP-VTON+, a more advanced virtual try-on framework dedicated to maintaining clothing shape and texture integrity during image-based try-ons. This method remedies problems like distortion and loss of texture by leveraging higher-order geometric matching and warping. CP-VTON+ particularly shines at keeping the structural shape of clothes while fitting them organically onto the human pose. Through emphasis on texture consistency and shape preservation, this method supplies a solid basis for realistic and detailed virtual try-on technology, and its suitability is extreme for e-commerce and fashion fields.

III. PROPOSED MODEL

Enhanced CAT-DM improves upon the virtual try-on process by building the best aspects of both diffusion models and GAN-based models under a truncation-based acceleration strategy. While diffusion models provide excellent generative flexibility and preservation of structure, their iterative sampling process is computationally costly for real-time inference. To counteract this, Enhanced CAT-DM includes a pre-trained GAN-based virtual try-on model to produce an initial coarse try-on image, which is used as the initialization for the diffusion process. Rather than denoising from clean Gaussian noise, noise is selectively added to the GAN-generated image, reducing the number of diffusion steps needed by orders of magnitude while maintaining critical garment details. This blend of approaches assures that Enhanced CAT-DM has computational efficiency and generation quality all at once. The GAN model offers a quick and realistic preliminary prediction, and the diffusion model refines fine details, like fabric textures and garment alignment. Through the adaptability in the truncation step, the framework dynamically balances the efficiency of the GAN and the accuracy of the diffusion model, supporting more controllable and high-fidelity virtual try-on synthesis

3.1 Garment-Conditioned Diffusion Model

The Enhanced CAT-DM further enhances the GC-DM architecture for better controllability and efficiency in virtual try-on applications. The improvement adds advanced feature extraction techniques along with a more sophisticated Poisson blending method to ensure real garment representation and smooth image synthesis.

3.1.1. Optimized ControlNet Architecture

Conventional diffusion models are computationally intensive and hence limit their use in real-time applications. To overcome this, Enhanced CAT-DM enhances the integration of ControlNet using the following approaches:

Lightweight Parameter Updates: Rather than retraining the entire diffusion model, ControlNet parameters are updated, thus saving GPU memory and computational power.

Extended Control Conditions: The model includes garment- agnostic representations like dense pose, pose maps, and segmentation masks to provide more control over garment alignment and fitting.

Refined Skip-Connections: By incorporating control vectors into skip-connections and the middle block of the U-Net, the model provides improved structural integrity in the output images.

The GC-DM model keeps the Pre-trained Base Encoder (PBE) with frozen parameters, utilizing its generative ability while improving local garment control. In the forward process, noise is added step by step to the person image, and the ControlNet generates control vectors that direct the reverse diffusion process to produce high-quality try-on images.

3.1.2. Improved Garment Feature Extraction

Maintaining garment patterns and textures accurately is an important task in virtual try-on work. Traditional diffusion models based on CLIP-based feature extraction usually cannot recover the fine-grained garment details. Improved CAT-DM resolves this by:

Substituting CLIP with DINO-V2: DINO-V2 allows patch- level feature representations to be richer, which lets the model recover finer garment structures.

Integrating Fully Connected Encoding Layers: Garment features are fed into a fully connected (FC) layer so that they can be made compatible with the cross-attention mechanism of the U-Net.

Cross-Attention Mechanism for Garment Integration: Garment features are inserted into the U-Net decoder via cross-attention so that the generated garment and the reference image are better consistent.

3.1.3 Optimized Poisson Blending for Seamless Integration

In conventional virtual try-on techniques, artifacts and boundary inconsistencies tend to be visible when overlaying generated clothing onto the original image. Improved CAT-DM enhances the Poisson blending method for seamless garment integration by the following improvements:

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025
DOI: 10.17148/IJARCCE.2025.14246

Adaptive Region Selection: The model selectively uses Poisson blending on regions according to garment edges, allowing for a natural blend with the original image.
Gradient-Based Boundary Refinement: A more refined gradient constraint guarantees that the boundary between the synthetic garment and the source image is still imperceptible.
Mask-Aware Blending: The blending strength is adaptively controlled based on garment occlusions and fabric types, maintaining realistic shading and texture continuity.

3.1.4 Computational Efficiency and Sampling Optimization

CAT-DM is optimized for computational efficiency by minimizing the cost of diffusion sampling with the following strategies:
GAN-Assisted Initialization: The pre-trained GAN-based model is utilized to create a coarse virtual try-on image, which acts as the starting point for the reverse diffusion process.
Noise-Adaptive Sampling: Rather than beginning with pure Gaussian noise, the model adds noise to the image generated by the GAN, which decreases the number of diffusion steps drastically.
Truncation Step (T_{trunc}): The truncation step determines the trade-off between GAN realism and diffusion flexibility, enabling adaptive trade-offs between generation speed and image quality.

3.2 Truncation based acceleration

Diffusion models usually need many sampling steps to produce high-quality images, so real-time applications are computationally expensive. Although techniques such as DDIM achieve efficient sampling, they are still required to use multiple iterations, and diffusion models tend to perform poorly in generating precise patterns and text on apparel when compared to GANs. To address these challenges, Enhanced CAT-DM utilizes a truncation-based acceleration strategy, which uses a pre-trained GAN as an initializer together with a fine-tuned diffusion process for high-quality image creation.
Rather than starting the reverse diffusion process from pure Gaussian noise, the model initially creates an initial try-on image with the help of a pre-trained GAN model. Controlled noise is added to create an implicit distribution at a truncated step (T_{trunc}), which acts as the initial point for the reverse diffusion chain. This helps the model avoid the initial diffusion steps, keeping the computational overhead very low while ensuring output quality.
The noisier image at T_{trunc} is then incrementally improved upon by the GC-DM framework, preserving structural details and garment accuracy. In contrast to traditional approaches, Enhanced CAT-DM utilizes DDIMs as samplers, minimizing denoising steps necessary while allowing for high-quality image synthesis. Rather than equally sampling throughout the full diffusion process, the model selectively improves over a truncated extent, balancing efficiency and image quality.
The truncation step (T_{trunc}) serves as a dynamic control mechanism, balancing the GAN and the GC-DM influence. A larger T_{trunc} yields more refinement by GC-DM, maintaining detailed garment textures and structure, whereas a smaller T_{trunc} maintains more features from the GAN output, providing quicker synthesis. This adaptive strategy maximizes computational efficiency with the retention of garment details, rendering Enhanced CAT-DM highly suitable for real-time virtual try-on applications.

3.3 Poisson Blending for Smooth Image Merging

Latent Diffusion Models (LDMs) employ pre-trained autoencoders to project images into a latent space, in essence minimizing computational complexity. Yet, such a projection has the potential to sacrifice pixel-level accuracy in reconstruction, especially in intricate regions such as faces, to produce discernible differences from the original image.
In order to obtain artifact-free fusion between the generated virtual try-on image and the original person image, Poisson blending is used. The method preserves the non-garment areas unaffected while blending the generated garment with the original image in a smooth manner. Instead of simply placing the generated garment on top of the input image and hence generating visible seams and discontinuities at the boundary, Poisson blending preserves smooth color and texture transitions.
In virtual try-on software, the objective is to substitute the clothing while maintaining the surrounding information. A naive method, for example, directly blending the input and generated images with a mask, tends to produce visible boundary artifacts. Through the modification of Poisson blending, these artifacts are reduced, resulting in a more natural and visually consistent outcome. This technique successfully removes stitching traces and increases the overall realism of the virtual try-on image.

3.4 DINO-V2: High-Quality Self-Supervised Feature Extraction

DINO-V2 is a cutting-edge self-supervised learning model that specializes in high-quality visual feature extraction. An improvement over its ancestor, DINO (self-Distillation with NO labels), it makes a number of improvements to make features more robust and accurate.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025
DOI: 10.17148/IJARCCE.2025.14246

In contrast to conventional supervised models, DINO-V2 extracts meaningful representations from images without needing labeled data, which makes it especially useful for applications where labeled datasets are limited or costly. One of its strengths is its capacity to extract global and local (patch-level) features, which allows it to capture subtle details in images. This feature is important in applications like virtual try-on, where fine-grained garment texture and pattern should be preserved. Architectural advancements in DINO-V2 increase the stability of extracted features so that they are robust in different datasets and tasks.

Its flexibility makes it suitable to be used for a broad set of computer vision tasks, ranging from image classification to object detection and segmentation. DINO-V2 also blends well with other models to enhance their performance. For example, in the CAT-DM project, it acts as the basic feature extractor of garment images with pixel-level information that enhances accuracy and realism when creating virtual try-on. Benchmark tests prove that DINO-V2 performs better consistently than both supervised and self-supervised models and is a worthy resource for high-quality image generation and editing. Being capable of extracting both overall structural patterns and minute visual details, DINO-V2 is a robust solution for intricate computer vision tasks with high precision and control requirements.

3.5 ControlNet in CAT-DM

ControlNet is essential in improving the controllability of the diffusion process in CAT-DM to allow for accurate garment alignment and realistic virtual try-on outcomes. Conventional diffusion models have difficulty with structural consistency, which tends to result in misaligned garments or misplaced details. With the addition of ControlNet, CAT-DM overcomes these issues through the addition of extra control signals, including dense pose, segmentation masks, and pose maps, to the diffusion setup.

In contrast to training the entire diffusion model, CAT-DM only optimizes the ControlNet parameters, which saves huge computational overhead but preserves high-quality outputs. ControlNet is incorporated into the U-Net backbone, where control vectors are inserted at various points, such as skip connections and the middle block, to reinforce structure preservation at every stage of the denoising process. This focused method ensures that the resulting garment is exactly proportional to the individual's body shape and posture and maintains key details like texture and pattern uniformity. Through the use of ControlNet, CAT-DM gains fine-grained control of the virtual try-on process and can generate realistic and smooth garment synthesis. Adding garment-agnostic control signals boosts flexibility, permitting the model to accommodate various styles of clothing and body poses. This innovation not only enhances generation accuracy but also renders real-time virtual try-on applications more computationally efficient and feasible.

IV. IMPLEMENTATION

The Enhanced CAT-DM project is organized into various main stages in order to obtain high-fidelity virtual try-on images in an efficient manner. The first stage of the process involves the input of a person image and target garment image. The heart of the project is the Garment-Conditioned Diffusion Model (GC-DM) that improves the controllability of the diffusion model for virtual try-on. GC-DM combines ControlNet with a pre-trained PBE model, freezing PBE's parameters to retain its generative ability while enhancing feature extraction from garment images. ControlNet produces control vectors from inputs such as noisy images, time steps, masks, and garment images, which are used to enhance controllability in PBE. Poisson blending is utilized to blend the original person image with the try-on image generated, so that regions outside the garment area are left unchanged.

For speeding up the process, this methodology adopts a truncation-based acceleration strategy. Rather than beginning with Gaussian noise, a pre-trained GAN-based model produces an initial try-on image, to which noise is added to form an implicit distribution. This provides the beginning for the reverse denoising operation, substantially lessening the steps of sampling that are needed. The output is a high-quality virtual try-on image that can realistically show the individual wearing the target clothing with retained details and fewer artifacts in fewer steps of sampling compared to conventional diffusion-based approaches. It synthesizes the ability of GANs and the robustness of diffusion models for a cutting-edge solution to the virtual try-on problem.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025

DOI: 10.17148/IJARCCE.2025.14246

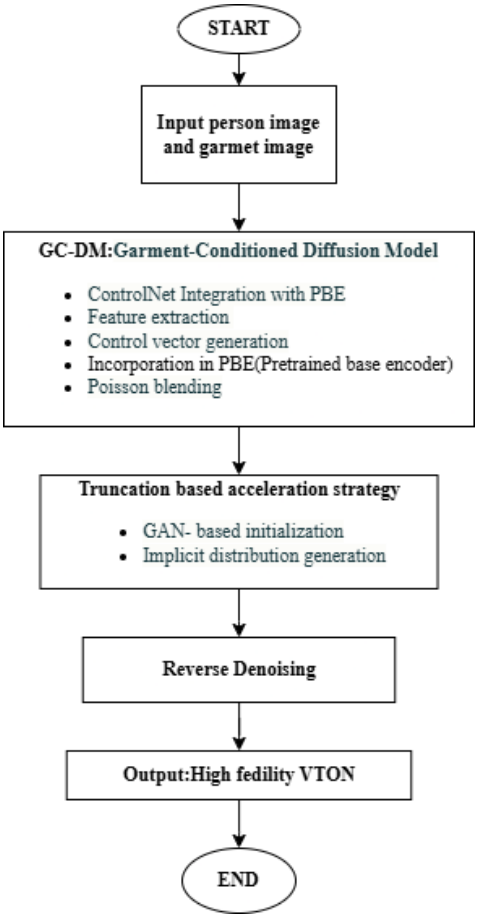


Figure 1: Flow of model implementation

The dataset includes a number of important elements required for complex virtual try-on operations, such as:

1. Person Image: High-quality images of people, which are the foundation for virtual try-on applications.
2. Dense pose: High-quality pose estimations that offer rich body surface information, which is important for precise garment fitting.
3. Parse-agnator: Segmentation masks that isolate the person from the background and other objects, allowing for accurate garment placement.
4. v3-agnator: More advanced segmentation and parsing capabilities that provide better accuracy and precision for intricate virtual try-on cases.
5. Open Pose: Skeletal pose predictions that detect the key points and joints of the body, which help with realistic draping and fitting of garments.
6. Agnostic: More abstract representations of the person image, usually discarding particular clothing elements to concentrate on the body form, to make it easier to overlay and manipulate the garment.

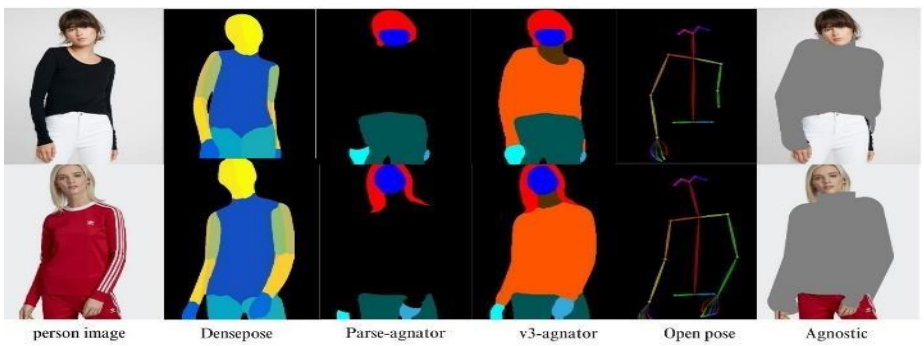


Figure 2: Sample Dataset

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025

DOI: 10.17148/IJARCCE.2025.14246

This complete data set allows for a strong basis for constructing and testing virtual try-on models in order to deliver high-quality, realistic outcomes.

V. RESULTS AND DISCUSSION

Table 1: Comparison with other models

METHOD	LPIPS	SSIM	CLIP-I
HR-VITON	0.330	0.741	0.701
LaDI-VTON	0.303	0.768	0.819
DCI-VTON	0.283	0.735	0.752
Stable-VITON	0.260	0.736	0.836
IDM-VTON	0.164	0.795	0.901
CAT-DM	0.0803	0.877	0.992
GC-DM	0.0988	0.862	0.741
Enhanced CAT-DM ***(ours)***	0.0801	0.877	0.992

This table is a comparison of different virtual try-on approaches with respect to three important metrics: LPIPS (Learned Perceptual Image Patch Similarity), SSIM (Structural Similarity Index Measure), and CLIP-I (CLIP Image Similarity). A description of the metrics and each method's performance follows:

LPIPS: This is the metric that assesses the perceptual similarity of two images. Lower values imply higher perceptual quality, which means the resulting image is closer to the reference image.

SSIM: This measurement evaluates the similarity of structure among images. A higher value indicates improved structural preservation, i.e., the resulting image preserves the structure of the original image better.

CLIP-I: This measurement compares the similarity between the resulting image and the reference image based on the CLIP model. A higher value indicates improved semantic correspondence and image quality.

Our approach, Enhanced CAT-DM, is the best-performing model in this experiment. It produces the best LPIPS value, demonstrating the highest perceived quality and correspondence with the comparison images. Its SSIM metric is also best, with clear evidence of effective structure preservation. Moreover, it has the largest CLIP-I score, proving higher semantic congruity and picture quality. These findings demonstrate the superiority of our method in producing high-quality virtual try-on images that are perceptually and structurally correct, and semantically well- aligned with reference images.

The Enhanced CAT-DM not only surpasses other state-of-the-art approaches but also establishes a new standard for virtual try-on tasks. This achievement is a reflection of the novel methodologies and careful optimization used in our project, and it is a top solution in the virtual try-on field. And following was the graphical representation of the above table.

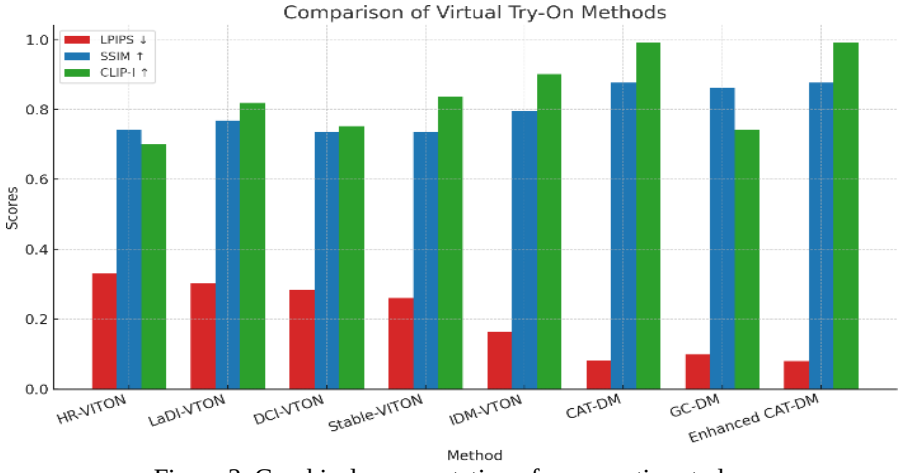


Figure 3: Graphical representation of comparative study

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025

DOI: 10.17148/IJARCCE.2025.14246

Screenshots of our implementation are

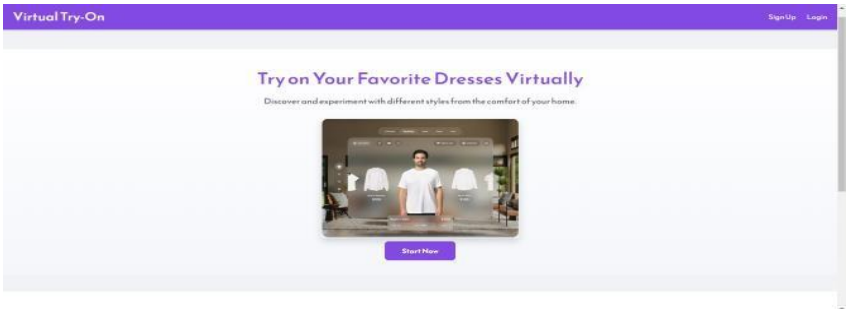


Figure 4: Home page of our implementation

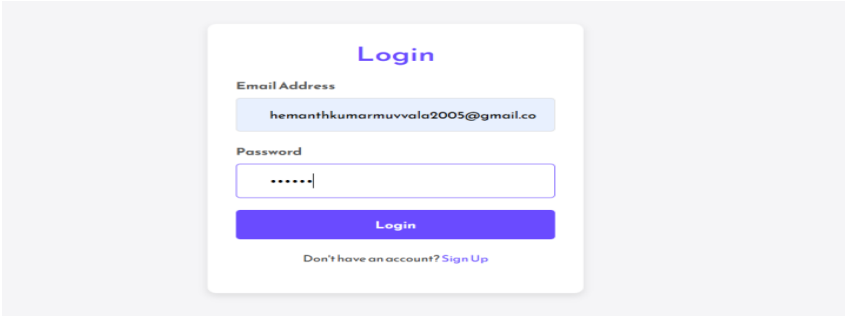
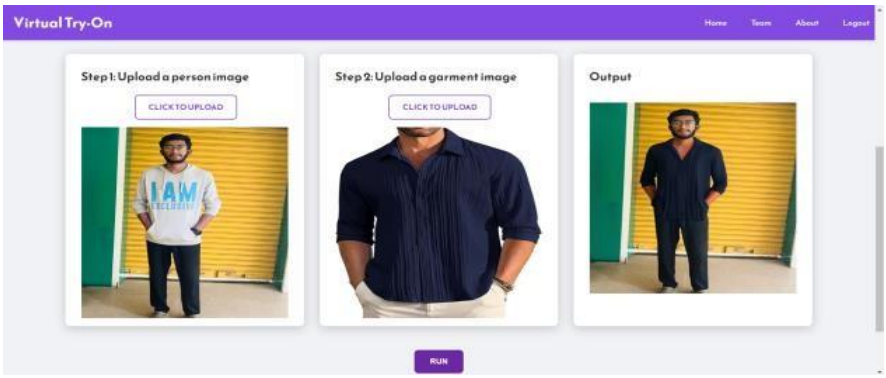


Figure 5: Login page of our implementation



Figur6: Synthesized Virtual Try-On Result 1

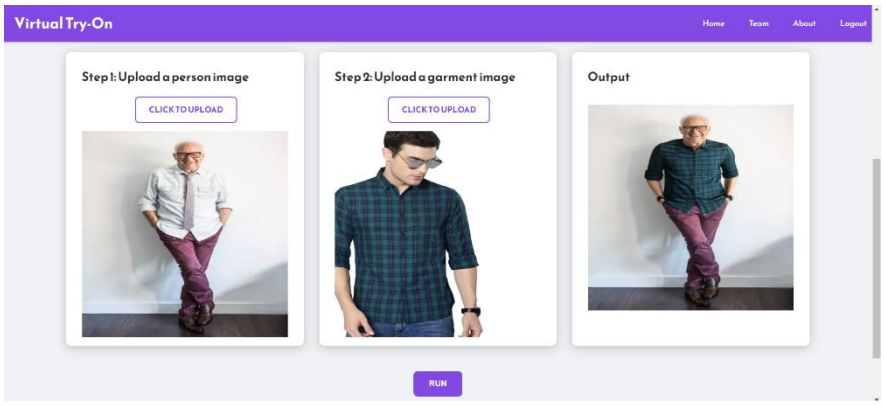


Figure 7: Synthesized Virtual Try-On Result 2

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025
DOI: 10.17148/IJARCCE.2025.14246

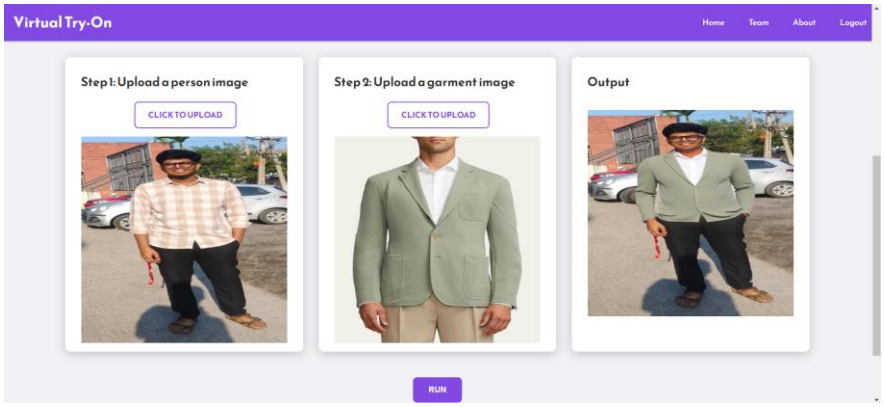


Figure 8: Synthesized Virtual Try-On Result 3

VI. CONCLUSION

The Advanced CAT-DM model greatly improves virtual try-on technology through better controllability, generation efficiency, and image quality. Based on the Garment-Conditioned Diffusion Model (GC-DM), and incorporating superior elements such as DINO-V2 and ControlNet, the model attains pixel-level garment reconstruction with outstanding accuracy. Texture preservation is enhanced by the incorporation of DINO-V2, and ControlNet provides fine-grained spatial control, leading to more natural and properly aligned try-on results. To maximize processing effectiveness, architecture utilizes a truncation-based acceleration technique in combination with GAN-aided initialization, which significantly diminishes the number of diffusion steps needed without affecting image quality. Such a combined method provides faster generation times without sacrificing detailed garment and body features. On top of that, the use of Optimized Poisson Blending contributes to realism by retaining facial and background details yet discarding artifacts from visibility, rendering a refined and professional visual output. Extensive testing with perceptual and semantic measures like LPIPS, SSIM, and CLIP-I validated that Enhanced CAT-DM is consistently better than earlier models in visual realism and computational efficiency. Through the accomplishment of an efficient balance between computational cost and accuracy, the Enhanced CAT-DM framework is a solid and pragmatic solution for real-world applications, especially in e-commerce, personalized fashion suggestions, and immersive virtual try-on experiences.

REFERENCES

[1]. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

[2]. Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, L. S. (2018). Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7543-7552).

[3]. Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3836-3847).

[4]. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600-612.

[5]. Bai, S., Zhou, H., Li, Z., Zhou, C., & Yang, H. (2022, October). Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision* (pp. 409-425). Cham: Springer Nature Switzerland.

[6]. Choi, S., Park, S., Lee, M., & Choo, J. (2021). Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14131-14140)

[7]. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

[8]. Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586-595).

[9]. Pérez, P., Gangnet, M., & Blake, A. (2023). Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2* (pp. 577-582).

[10]. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., & Luo, P. (2021). Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8485-8493).

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025

DOI: 10.17148/IJARCCE.2025.14246

[11]. Yang, H., Yu, X., & Liu, Z. (2022). Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3460-3469).

[12]. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., ... & Wen, F. (2023). Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18381-18391).

[13]. Li, Z., Wei, P., Yin, X., Ma, Z., & Kot, A. C. (2023). Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 22788-22797).

[14]. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., & Cucchiara, R. (2022). Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2231-2235).

[15]. Minar, M. R., Tuan, T. T., Ahn, H., Rosin, P., & Lai, Y. K. (2020, June). Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR workshops* (Vol. 3, pp. 10-14).

[16]. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., ... & Liang, X. (2023). Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 23550-23559).

[17]. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

[18]. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., ... & Kemelmacher-Shlizerman, I. (2023). Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4606-4615).

[19]. Zeng, J., Song, D., Nie, W., Tian, H., Wang, T., & Liu, A. A. (2024). CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8372-8382).

[20]. Ian, J. (2014). Goodfellow, jean pouget-abadie, mehdi mirza, bing xu, david warde-farley, sherjil ozair, aaron courville, yoshua bengio. generative adversarial networks. *Advances in Neural Information Processing Systems*, *27*, 8-13.

[21]. Chen, C. Y., Chen, Y. C., Shuai, H. H., & Cheng, W. H. (2023). Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7513-7522).

