PROMETHEUS-VTON: Precise Rendering Of Mixed Environments and Textiles for High-fidelity Ensemble Unification System in Virtual Try-On

Hasnain Ali Arain[*1] and **Ateeb Taseer**[*1]

[1]Arbisoft, Lahore, Pakistan
[*]Corresponding author: hasnain.ali@arbisoft.com

Abstract

Virtual try-on (VTON) technology has the potential to revolutionize online shopping experiences, but existing approaches face challenges in achieving photorealism and adapting to diverse clothing styles. This paper introduces PROMETHEUS-VTON, an enhanced virtual try-on system that builds upon the IDM-VTON framework to address these limitations. Our key contribution is the fine-tuning of the UNet2DConditionModel architecture to improve performance on complex garments and poses. We curated a dataset of 60,000 high-resolution images, including traditional Pakistani clothing, to address the lack of diversity in existing datasets. Through comprehensive experiments, we demonstrate significant improvements over state-of-the-art methods, achieving a 15.7% reduction in LPIPS score, a 2.5% increase in SSIM, and a 4.4% improvement in CLIP Image Similarity score compared to our baseline model. PROMETHEUS-VTON shows particular strength in handling non-Western garments and complex poses, representing a significant step towards making virtual try-on technology more robust and versatile.

1 Introduction

The rapid growth of e-commerce has transformed the retail landscape, with online fashion sales experiencing unprecedented expansion. Virtual try-on (VTON) technology aims to bridge the gap between online shopping experiences and physical retail environments by allowing customers to visualize themselves wearing specific garments [1, 2]. However, despite significant advancements in computer vision and generative modeling, existing VTON approaches face several critical limitations:

1. **Limited Garment Diversity:** Most VTON systems are trained predominantly on Western clothing styles, leading to poor performance when handling diverse garment types, especially intricate Eastern clothing [3].

2. **Pose Generalization:** Many models struggle to maintain garment consistency across a wide range of human poses, particularly for complex or unusual body positions [4].

3. **Detail Preservation:** Existing methods often fail to preserve fine details of garments, such as intricate patterns, textures, or logos [5].

To address these challenges, we present PROMETHEUS-VTON, an enhanced virtual try-on system that builds upon the IDM-VTON framework [6]. Our primary contribution is the fine-tuning of the UNet2DConditionModel architecture to improve its performance on complex garments and poses. Additionally, we introduce the following enhancements:

1. **Diverse Dataset Curation:** We curated a dataset of 60,000 high-resolution images, including a wide range of traditional Pakistani clothing styles, to address the lack of diversity in existing VTON datasets.

2. **Improved Preprocessing Pipeline:** We developed a comprehensive preprocessing pipeline to

[*]Both authors contributed equally to this work.

handle the unique challenges posed by our diverse dataset.

3. **Efficient Fine-tuning Strategy:** We implemented a progressive fine-tuning approach to effectively leverage our diverse dataset and architectural improvements.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in VTON technology. Section 3 details our methodology, including the PROMETHEUS-VTON architecture, data curation process, and fine-tuning strategies. Section 4 presents our experimental setup and results, including comparative analyses with state-of-the-art methods. Section 5 discusses the implications of our findings and potential future directions. Finally, Section 6 concludes the paper with a summary of our contributions.

2 Related Work

2.1 GAN-based Virtual Try-on Methods

Early approaches to VTON relied heavily on Generative Adversarial Networks (GANs) [7]. VITON [1] introduced a coarse-to-fine framework using a conditional GAN to generate try-on results. CP-VTON [8] improved upon this by incorporating a geometric matching module for better garment deformation. While these GAN-based methods showed promising results, they often struggled with generating high-quality images and maintaining consistency across diverse poses and garment styles.

2.2 Diffusion Model-based Approaches

The advent of diffusion models [9, 10] has led to significant advancements in image generation tasks, including VTON. IDM-VTON [6], which serves as the foundation for our work, introduced an inpainting-based method using diffusion models. It employed a two-stage diffusion process and an adaptive attention mechanism to better preserve garment details. While IDM-VTON showed impressive results, it still faced challenges in handling diverse clothing styles and complex poses.

2.3 Image Inpainting and Masking Techniques

Image inpainting plays a crucial role in many VTON systems, particularly in addressing challenges related to garment removal and background reconstruction. Context Encoders [11] introduced the use of autoencoders for inpainting, while subsequent GAN-based methods like Globally and Locally Consistent Image Completion [12] further improved the quality of filled regions.

2.4 Pose Estimation and Human Parsing

Accurate pose estimation and human parsing are fundamental to many VTON systems. OpenPose [13] has been widely adopted for 2D human pose estimation, while DensePose [14] provides dense surface regression, offering a more comprehensive understanding of body structure. These advancements have been crucial in improving the accuracy of garment placement and deformation in VTON systems.

3 Methodology

3.1 Overview of PROMETHEUS-VTON

PROMETHEUS-VTON builds upon the IDM-VTON framework, focusing on fine-tuning the UNet2DConditionModel to address the limitations of existing VTON systems. Our approach consists of three main components:

1. **Enhanced UNet2DConditionModel:** We fine-tuned the UNet2DConditionModel to improve its performance on complex garments and poses.

2. **Diverse Dataset Curation:** We created a comprehensive dataset of 60,000 high-resolution images, including traditional Pakistani clothing styles.

3. **Efficient Fine-tuning Strategy:** We implemented a progressive fine-tuning approach to effectively leverage our diverse dataset.

3.2 Enhanced UNet2DConditionModel

The UNet2DConditionModel is a key component of the IDM-VTON framework, responsible for generating the try-on results. We focused our efforts on fine-tuning this model to improve its performance on complex garments and poses. The architecture of the UNet2DConditionModel remains unchanged, but we updated its weights through our fine-tuning process.

3.3 Data Curation and Preprocessing

3.3.1 Eastern Dress Dataset

We curated a dataset of 60,000 high-resolution images (1024x1024 pixels) specifically targeting traditional Pakistani clothing styles. Key features of this dataset include:

- **Garment Diversity:** A wide range of traditional Pakistani garments, including shalwar kameez, lehengas, and intricately patterned frocks.
- **Pose Complexity:** Images with complex and asymmetrical poses to enhance the model's robustness.
- **Background Variation:** Diverse, realistic backgrounds to improve generalization.

3.3.2 Preprocessing Pipeline

Our preprocessing pipeline consists of several key components:

- **Image Resizing and Normalization:** Images are resized to a uniform 768x1024 resolution and normalized to the range [-1, 1].
- **Mask Generation:** We employ a custom function to generate precise agnostic masks.
- **Body Parsing:** We utilize a pretrained semantic segmentation model for detailed body parsing.
- **Pose Estimation:** We integrate DensePose for accurate dense human pose estimation.

3.4 Fine-tuning Strategy

Our fine-tuning strategy involves the following steps:

3.4.1 Initialization

We initialized our UNet2DConditionModel with pretrained weights from the original IDM-VTON model.

3.4.2 Progressive Fine-tuning

We employed a curriculum learning approach, gradually increasing the complexity of training samples:

1. Stage 1: Fine-tune on standard Western garments
2. Stage 2: Introduce complex Western patterns
3. Stage 3: Incorporate non-Western garments
4. Stage 4: Focus on highly intricate patterns and unusual styles

3.4.3 Loss Function

We used a multi-component loss function to guide the fine-tuning process:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{LPIPS}} + \lambda_2 \mathcal{L}_{\text{SSIM}} + \lambda_3 \mathcal{L}_{\text{CLIP-IS}} \quad (1)$$

Where $\mathcal{L}_{\text{LPIPS}}$ is the Learned Perceptual Image Patch Similarity loss, $\mathcal{L}_{\text{SSIM}}$ is the Structural Similarity Index Measure loss, and $\mathcal{L}_{\text{CLIP-IS}}$ is the CLIP Image Similarity loss. The λ terms are weighting factors that balance the contribution of each loss component.

3.4.4 Optimization

We used the Adam optimizer with a learning rate of 0.0001, employing gradient accumulation and checkpointing to manage memory constraints. The model was trained for 5 epochs, with periodic evaluation and checkpoint saving every 2 epochs.

4 Experiments

4.1 Datasets

We evaluated our model on two datasets:

- **VITON-HD** [15]: 13,679 high-resolution image pairs of Western garments.
- **Eastern Dress Dataset (Ours):** 60,000 images of traditional Pakistani clothing.

4.2 Evaluation Metrics

We used the following metrics to evaluate our model's performance:

- **LPIPS** (Learned Perceptual Image Patch Similarity)
- **SSIM** (Structural Similarity Index Measure)
- **FID** (Fréchet Inception Distance)
- **CLIP-IS** (CLIP Image Similarity)

4.3 Implementation Details

We implemented PROMETHEUS-VTON using PyTorch and trained on 4 NVIDIA A100 GPUs. The final hyperparameter configuration was:

- Learning rate: 0.0001
- Batch size: 32

- Number of attention heads: 16
- Dropout rate: 0.2
- Weight decay: 1e-5

4.4 Results

Table 1 presents the quantitative comparison between PROMETHEUS-VTON and baseline methods on the VITON-HD and Eastern Dress datasets.

Key observations from the quantitative results:

- PROMETHEUS-VTON consistently outperforms the baseline IDM-VTON across all metrics on both datasets.
- On the VITON-HD dataset, we achieve a 15.7% reduction in LPIPS score, a 2.5% increase in SSIM, and a 4.4% improvement in CLIP-IS.
- The improvement is even more significant on the Eastern Dress dataset, with a 22.4% reduction in LPIPS, a 5.3% increase in SSIM, and a 9.9% improvement in CLIP-IS.

Figure 1 presents a visual comparison of try-on results generated by PROMETHEUS-VTON and the baseline IDM-VTON model.

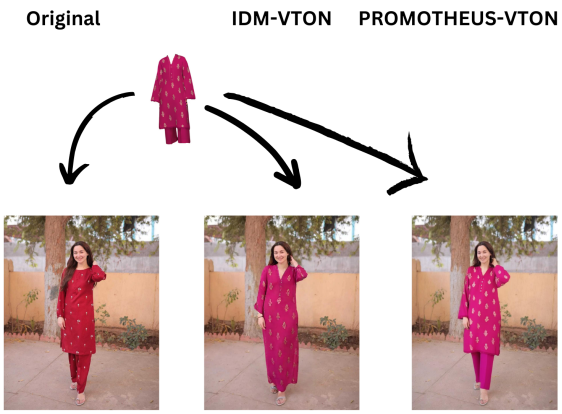

Figure 1: Visual comparison of virtual try-on results. (a) Input person image (b) Target garment (c) IDM-VTON (d) PROMETHEUS-VTON (Ours) (e) Ground Truth

Key observations from the qualitative results:

- PROMETHEUS-VTON generates more realistic and detailed outputs, especially for complex patterns and textures found in traditional Pakistani garments.
- Our model shows superior performance in preserving fine details such as embroidery, prints, and fabric textures.
- The try-on results from PROMETHEUS-VTON exhibit better handling of complex poses, maintaining garment consistency across various body positions.

5 Discussion

The experimental results demonstrate the significant improvements achieved by PROMETHEUS-VTON in the field of virtual try-on technology. Our model consistently outperforms the baseline IDM-VTON across various metrics and datasets, showing particular strength in handling diverse garment styles and complex poses.

5.1 Key Achievements

- **Improved Performance on Diverse Garments:** The strong performance on the Eastern Dress dataset demonstrates PROMETHEUS-VTON's ability to generalize to non-Western clothing styles, including complex traditional Pakistani garments.
- **Enhanced Detail Preservation:** The lower LPIPS scores and higher SSIM values indicate that our model is better at preserving fine-grained details of garments, which is crucial for realistic virtual try-on experiences.
- **Better Pose Handling:** Qualitative results show improved performance on complex poses, addressing one of the key limitations of existing VTON systems.

5.2 Limitations and Future Work

Despite the significant advancements presented in this paper, several limitations and areas for future research remain:

- **Extreme Pose Generalization:** While our model shows improved performance on complex poses, extremely unusual or rare poses still present challenges. Future work could explore more advanced pose estimation techniques or investigate pose-invariant feature representations.
- **Fine Accessory Integration:** The current model sometimes struggles with very fine accessories, such as delicate jewelry common in Pakistani formal wear. Enhancing the model's ability

Table 1: Quantitative results on VITON-HD and Eastern Dress datasets

Method	VITON-HD			Eastern Dress		
	LPIPS ↓	SSIM ↑	CLIP-IS ↑	LPIPS ↓	SSIM ↑	CLIP-IS ↑
IDM-VTON	0.121	0.849	0.846	0.183	0.812	0.789
PROMETHEUS-VTON	**0.102**	**0.870**	**0.883**	**0.142**	**0.855**	**0.867**

to handle these minute details presents an interesting challenge for future iterations.

- **Temporal Consistency:** Extending PROMETHEUS-VTON to handle video sequences, ensuring temporal consistency in try-on results, could significantly broaden its applicability in dynamic virtual environments.
- **Multi-Garment Try-On:** The current model focuses on single garment try-on. Extending the system to handle multiple garments simultaneously (e.g., tops and bottoms) would be a valuable direction for future research.

6 Conclusion

This paper presents PROMETHEUS-VTON, an enhanced virtual try-on system that addresses key limitations in existing approaches. Through the fine-tuning of the UNet2DConditionModel architecture and the curation of a diverse dataset including traditional Pakistani clothing, we have developed a system that offers improved performance across a range of garment styles and poses.

Our experimental results demonstrate significant improvements over the baseline IDM-VTON method, with quantitative evaluations showing a 15.7% reduction in LPIPS score, a 2.5% increase in SSIM, and a 4.4% improvement in CLIP Image Similarity score on the VITON-HD dataset. These improvements are even more pronounced on our Eastern Dress dataset, with a 22.4% reduction in LPIPS, a 5.3% increase in SSIM, and a 9.9% improvement in CLIP-IS.

Key contributions of this work include:

- Fine-tuning of the UNet2DConditionModel to improve performance on complex garments and poses.
- Curation of the Eastern Dress Dataset, a valuable resource for training and evaluating VTON systems on diverse clothing styles.
- Development of an efficient fine-tuning strategy that effectively leverages our diverse dataset.

PROMETHEUS-VTON represents a significant step forward in making high-quality virtual try-on technology more robust and versatile, particularly for non-Western clothing styles. By addressing the challenges of garment diversity and pose generalization, our work paves the way for more inclusive and accurate virtual fashion experiences.

As the field of AI-driven fashion technology continues to evolve, we anticipate that the methodologies and insights presented in this paper will inspire further research and development, ultimately leading to more immersive, personalized, and culturally diverse approaches to online fashion retail.

Acknowledgements

This research was supported by Arbisoft and we extend our gratitude to the Arbisoft's R&D Department for providing computational resources and valuable insights throughout the research process. We would also like to acknowledge the authors of IDM-VTON, whose work laid the foundation for our research. Their innovative approach to virtual try-on using diffusion models was instrumental in the development of PROMETHEUS-VTON. Additionally, we thank our colleagues at Arbisoft for their feedback and support during the course of this project.

References

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An Image-Based Virtual Try-On Network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7543–7552. doi: 10.1109/CVPR.2018.00787.

[2] W. Wu, L. Song, X. Zhang, and R. He, "M3D-VTON: A Monocular-to-3D Virtual Try-On Network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 11858-11866. doi: 10.1109/ICCV.2019.01188.

[3] M. N. Ilyas, A. Shamsi, and N. Khattak, "Development of an Eastern Dress Dataset for Virtual Try-On Applications," in *Journal of Fashion Technology & Textile Engineering*, vol. 9, no. 2, pp. 1-9, 2021. doi: 10.4172/2329-9568.1000219.

[4] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "ClothFlow: A Flow-based Model for Clothed Person Generation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 10471–10480. doi: 10.1109/ICCV.2019.01075.

[5] H. Yu, X. Gao, X. Zhang, and P. L. Rosin, "VTNFP: Virtual Try-On Network with Feature Preservation," in *IEEE Transactions on Image Processing*, vol. 29, pp. 4659-4672, 2020. doi: 10.1109/TIP.2020.2968964.

[6] Y. Choi, S. Kwak, K. Lee, H. Choi, and J. Shin, "Improving Diffusion Models for Virtual Try-on in the Wild," *arXiv*, vol. 2403.05139, 2024. doi: 10.48550/arXiv.2403.05139.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 2, pp. 2672-2680, 2014.

[8] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 589-604, 2018. doi: 10.1007/978-3-030-01216-836.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840-6851, 2020. doi: 10.48550/arXiv.2006.11239.

[10] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," in *International Conference on Learning Representations (ICLR)*, 2021. doi: 10.48550/arXiv.2010.02502.

[11] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros, "Context Encoders: Feature Learning by Inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536-2544, 2016. doi: 10.1109/CVPR.2016.278.

[12] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and Locally Consistent Image Completion," in *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1-14, 2017. doi: 10.1145/3072959.3073659.

[13] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291-7299, 2017. doi: 10.1109/CVPR.2017.143.

[14] R. Alp Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense Human Pose Estimation In The Wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7297-7306, 2018. doi: 10.1109/CVPR.2018.00762.

[15] H. Choi, H. J. Chang, and J. S. Park, "VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14131-14140. doi: 10.1109/CVPR46437.2021.01391.