TryFit: A Deep Learning-Based Virtual Try-On System for Modest and Regional Fashion

Genyveyav Raafat Louka
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams Univercity
Cairo, Egypt
genyveyavraafat@gmail.com

Hassnaa Hassan Saied
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams Univercity
Cairo, Egypt
hassnaashanan585@gmail.com

Habiba Mohammed Yahia
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams Univercity
Cairo, Egypt
habibamohammedyahia@gmail.com

Toqa Ossama Ali
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams Univercity
Cairo, Egypt
toqaosama673@gmail.com

Monica Adel Lotfy
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams Univercity
Cairo, Egypt
monicaadel543@gmail.com

Maria George Kamel
Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams Univercity
Cairo, Egypt
mariageorgekamel@gmail.com

Prof. Dr Abeer Mahmoud
Professor ,Computer Science Department
Faculty of Computer and Information Sciences, Ain Shams Univercity

T.A. Mohamed Essam
Tecching Assistant
Bioinformatics Department
Faculty of Computer and Information Sciences, Ain Shams Univercity

***Abstract*—**In Egypt's growing online fashion market, modesty and cultural sensitivity are often overlooked in virtual try-on (VTON) systems. TryFit addresses this challenge by providing a culturally aware, hijab-supportive VTON solution tailored for Egyptian users. The system allows users to upload their images and virtually try on upper, lower, and full outfits, including hijabs, while preserving realism and modesty. It is powered by a modular deep learning pipeline that integrates SCHP for human parsing, OpenPose for pose estimation, TPS warping, EMASC refinement, and a LaDi-VTON-based latent diffusion model for image synthesis. The platform is deployed with a Flutter-based mobile interface and Firebase backend, offering secure access, product management, and result downloads. TryFit enhances user confidence in digital fashion experiences while supporting local brands and inclusive design practices.

Keywords—Virtual Try-On, Latent Diffusion, Hijab Try-On, Egyptian Brands

I. INTRODUCTION

The rise of online fashion retail has introduced both convenience and complexity into the shopping experience. One of the most persistent challenges faced by online shoppers is the inability to visualize how clothing will look and fit on their own bodies before purchasing. This issue often results in low purchase confidence, high return rates, and customer dissatisfaction — particularly in culturally specific markets like Egypt, where modesty, local fashion identity, and privacy are essential considerations.

The **TryFit** project addresses these gaps through a culturally aware **Virtual Try-On (VTON)** system tailored to the needs of Egyptian users, including veiled women. The system enables users to upload their image, select garments, and receive realistic try-on results — all within a user-friendly mobile application. It supports not only standard upper, lower, and full outfits but also hijab styles and local fashion brands that are often underrepresented in existing platforms.

By combining modern deep learning techniques — such as pose estimation, image segmentation, and latent diffusion models — with an inclusive dataset, **TryFit** delivers personalized, private, and realistic visualizations that enhance the decision-making process in digital shopping. Moreover, the system features dual user roles: normal users who engage with try-on functionalities and providers who manage product content.

II. RELATED WORK

Over the past few years, significant research has been conducted in the field of VTON, with a strong focus on realism, pose alignment, and garment preservation. Despite these advancements, existing systems are often limited in terms of cultural diversity, particularly when it comes to modest fashion, hijab representation, and support for regional clothing preferences such as those prevalent in Egypt.

One of the most notable systems is **VITON-HD** [1], which generates high-resolution try-on results for upper-body garments using pose-guided warping and refinement stages. While effective in preserving clothing details and body alignment, VITON-HD is restricted to **upper garments only** and lacks full-body or lower-body support. Additionally, it does not consider veiled users or modest fashion requirements.

CAT-VTON [2] introduced a content-aware try-on mechanism that improves compatibility between the target clothing and the model by learning contextual relations. Although effective for generating cleaner try-on results, it is

trained on Western datasets, without addressing culturally specific garments such as **abayas**, **hijabs**, or **modest fashion**.

In contrast, **TryFit** distinguishes itself by integrating hijab-specific clothing items, supporting full outfits (upper, lower, and full-body), and providing a try-on experience tailored for **Egyptian users**, including veiled women. It also introduces support for **local Egyptian brands** and offers category-based browsing options, which are largely absent in previous works.

Through the use of culturally diverse data and practical mobile implementation, TryFit not only extends technical contributions but also addresses **inclusivity and usability in underserved communities**.

III. SYSTEM ARCHITECTURE

The system architecture is designed to automate the process of:
Generating realistic virtual try-ons for modest fashion (hijabs, abayas)
Aligning garments precisely with diverse body poses and shapes
Preserving cultural modesty during digital fittings
The architecture integrates specialized modules deployed via:
Flask API: Backend processing and model inference
Flutter: Cross-platform mobile interface
Firebase: Secure storage for user data, garment categories, and authentication

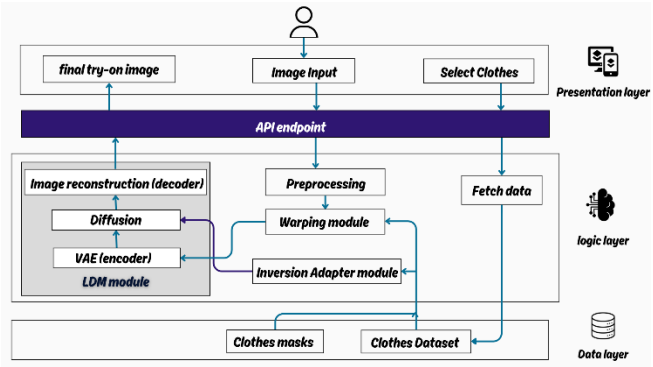

Figure 1 System Architecture

As shown in Figure 1, TryFit's modular architecture automates realistic virtual try-ons through a three-layer workflow, optimized for modest fashion:

1. Presentation Layer (Flutter UI)

User Interaction:
Upload full-body photos via gallery/camera.
Browse categorized clothing (Upper/Lower/Full) with hijab-friendly filters.
Output: Displays try-on results with download option.

2. Logic Layer (Flask API + AI Core)

Phase 1: Preprocessing

Human Parsing: SCHP segments body regions (hijab, skin) for modesty preservation.
Pose Estimation: OpenPose extracts 18 keypoints for anatomical alignment.
Cloth Masking: Grounded-SAM generates precise binary masks for garment isolation.

Phase 2: Garment Alignment

TPS Warping: Aligns clothing to user pose using keypoints and masks.
EMASC Refinement: Mask-aware skip connections enhance edges (e.g., loose sleeves, hijab draping).
CLIP Embeddings: V⋆ tokens encode garment semantics (texture, style).
Phase 3: Diffusion Synthesis
LaDI-VTON: Latent Diffusion Model (LDM) synthesizes outputs conditioned on:
Warped garments + masks
Pose maps + CLIP embeddings
VAE Decoder: Generates 1024×768px photorealistic images.

3. Data Layer (Firebase)
Storage: Clothing categories (Upper, Lower, Full) and user try-on histories.
Authentication: Secure login/profile management.

IV. RESULTS

This section presents the experimental evaluation of the TryFit system across multiple datasets. The goal was to assess image synthesis quality, cultural relevance, and the effectiveness of key components such as garment warping, refinement, and preprocessing. The evaluation included both quantitative metrics and qualitative observations across upper, lower, and full-body garment categories.

1) Experimental Setup

TryFit was tested on four datasets collected from real-world Egyptian fashion sources, covering veiled and unveiled models, a variety of garment types, and diverse poses and lighting conditions:

Dataset 1: 706 pairs of images (121 veiled)
Dataset 2: 586 pairs of images (10 veiled)
Dataset 3: 1274 pairs of images (130 veiled)
Dataset 4: 1912 pairs of images

Each configuration varied the number of refinement steps, warping complexity, and the inclusion of semantic preprocessing components such as face blurring and label-based masking. We evaluated the try-on results using:

Table I: Quantitative analysis for fine-tuning

Model	SSIM ↑	FID ↓	LPIPS ↓
CP-VTON	0.842	28.44	0.186
ACGPN	0.858	**12.96**	_
LaDi VTON	0.848	58.35	0.154
Ours	**0.868**	53.28	**0.140**

2) Quantitative Results

Table II summarizes the results across different configurations. Experiment 5, conducted on Dataset 3 with full preprocessing and hyperparameter tuning, achieved the best performance with an FID of **53.28** and an LPIPS of **0.140**.

Frechet Inception Distance (FID) ↓: Measures image realism (lower is better)
LPIPS (Learned Perceptual Image Patch Similarity) ↓: Measures perceptual similarity (lower is better)
Structural Similarity Index (SSIM) ↑: Measures structural and perceptual similarity between the generated and real image (higher is better)

TABLE II: EXPERIMENTAL RESULTS ACROSS DATASETS USING FID AND LPIPS METRICS

#	Dataset	EMASC steps	TPS epochs	refinement epochs	Text	descriptions	Hyper parameter tuning	FID ↓	LPIPS ↓
1	Dataset 1	2000	5	5	_	_	_	111.35	0.24
2	Dataset 1	3000	5	5	_	_	_	78.49	0.147
3	Dataset 2	3000	30	30	_	_	_	103.22	0.147
4	Dataset 3	5000	50	50	✓	_	_	53.577	0.142
5	Dataset 3	7000	50	50	✓	_	✓	53.283	0.140
6	Dataset 4	7000	75	75	✓	✓	✓	57.219	0.168
7	Dataset 4	7000	70	70	✓	✓	✓	59.249	0.169
8	Dataset 4	7000	70	70	✓	_	✓	57.62	0.168

The best performance was achieved in **Experiment 5** using **Dataset 3**, where **hyperparameter tuning** and **semantic garment encoding** via CLIP text embeddings led to improved output quality. This setup yielded the **lowest FID (53.283)** and **LPIPS (0.140)** scores, indicating high visual realism and strong perceptual similarity. The use of **V⋆ tokens** helped the model better represent garment style and texture, while tuning improved convergence and consistency.

3) Qualitative Observations

a) Full-Body Veiled Garments

The system successfully preserved the structure of hijabs, showing no noticeable distortions or unnatural blending with the background or face region. Long dresses were well-aligned with the body pose, maintaining consistent texture and flow from the upper body to the hem. This consistency

highlights the model's suitability for modest fashion, where full-body coverage and structural integrity are crucial.

b) Upper Garments (Blouses, Jackets)

For upper-body clothing such as blouses, shirts, and jackets, TPS-based warping achieved reliable alignment around the shoulders and arms. The EMASC refinement module further improved garment realism by enhancing sleeve connectivity and reducing common issues like disconnected cuffs or unnatural folds. The use of face blurring helped minimize distractions, bringing visual focus to the garments themselves.

c) Lower Garments (Pants, Skirts)

Lower garments, including pants and skirts, were rendered effectively when pose keypoints were accurately extracted. Skirts exhibited a realistic drape and preserved shape under typical body postures. Minor visual glitches, such as slight misalignments or texture warping, occurred in cases of occlusion (e.g., crossed legs), indicating that future improvements in pose estimation and garment masking would enhance overall quality.

4) Visual Comparisons

Side-by-side visual comparisons highlighted improvements at various stages:

Warped garments before and after EMASC refinement

Try-on images with and without face blurring

Figure 2: Visual Results

Category-specific results: veiled dresses, upper wear, and lower wear

These visualizations supported the quantitative findings, confirming that TryFit produces culturally respectful and realistic virtual try-on results.

LIMITATIONS

The model exhibits sensitivity to occlusions and pose estimation errors, occasionally resulting in misaligned or distorted garment outputs. Moreover, the scarcity of paired training data for hijab and other modest fashion garments limits the model's ability to generalize across different styles, fabrics, and draping variations.

5) Summary of Findings

The best performance was observed when the complete preprocessing pipeline was applied, especially in the case of **veiled full-body garments**. The integration of **EMASC** and **TPS modules** proved essential for achieving accurate garment alignment and maintaining edge coherence around complex regions such as sleeves, hijabs, and hemlines. These components contributed significantly to the natural appearance and structural integrity of the try-on results. The application of **face blurring** not only enhanced **user privacy** but also reduced visual distractions, allowing clearer evaluation of garment fit and texture. However, **minor limitations** persist, particularly in **lower-body alignment** during challenging poses such as crossed legs or side turns. These issues indicate the need for further enhancements in **pose estimation accuracy** and **occlusion handling** to improve consistency in such scenarios.

V. CONCLUSION

This paper presented **TryFit**, a culturally aware virtual try-on system specifically designed to meet the needs of **modest fashion users in Egypt**. The system leverages a powerful deep learning pipeline that includes **SCHP** for detailed human parsing, **OpenPose** for pose estimation, **TPS** transformation for non-rigid garment warping, **EMASC** for refinement of garment structure and alignment, and **LaDI-VTON** for realistic and high-resolution image generation in latent space. These components work together to produce **photorealistic, pose-consistent try-on outputs**, with accurate handling of **hijabs and full-body garments**. TryFit is deployed using a modular architecture that combines **Flutter** for mobile frontend development, **Flask** for the backend inference engine, and **Firebase** for secure data storage and real-time synchronization. This setup ensures **scalability, privacy, and user accessibility** across devices. By supporting modest fashion and region-specific garments, TryFit highlights the potential of AI-driven solutions to **enhance user trust, reduce return rates**, and **empower local fashion brands** through inclusive, culturally sensitive virtual try-on technology.

ACKNOWLEDGMENT

First and foremost, all praise and gratitude are due to Allah, whose boundless mercy and guidance empowered us throughout this journey.

We extend our deepest appreciation to:

Our **families**, especially our parents, for their unwavering support, sacrifices, and prayers.

Prof. Dr. Abeer Mahmoud and **T.A. Mohamed Essam** for their invaluable mentorship, critical feedback, and steadfast encouragement during every phase of this project.

Our **friends and colleagues** for their camaraderie during late-night debugging sessions and moments of doubt.

We also acknowledge the foundational work of open-source projects including **SCHP human parsing**[11], **OpenPose**[12], and **LaDI-VTON**[3], which enabled key components of our system.

REFERENCES

[1] S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization." Accessed: Jun. 11, 2025. https://openaccess.thecvf.com/content/CVPR2021/papers/Choi_VITON-HD_High-Resolution_Virtual_Try-On_via_Misalignment-Aware_Normalization_CVPR_2021_paper.pdf.

[2] Z. Chong *et al.*, "CatVTON: Concatenation Is All You Need for Virtual Try-On with Diffusion Models," *arXiv.org*, 2024. https://arxiv.org/abs/2407.15886.

[3] D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On," Oct. 2023, doi: https://doi.org/10.1145/3581783.3612137.

[4] Y. Choi, S. Kwak, K. Lee, H. Choi, and J. Shin, "Improving Diffusion Models for Authentic Virtual Try-on in the Wild," *Lecture Notes in Computer Science*, pp. 206–235, Oct. 2024, doi: https://doi.org/10.1007/978-3-031-73016-0_13.

[5] Y. Xu, T. Gu, W. Chen, and A. Chen, "OOTDiffusion: Outfitting Fusion Based Latent Diffusion for Controllable Virtual Try-On," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, pp. 8996–9004, Apr. 2025, doi: https://doi.org/10.1609/aaai.v39i9.32973.

[6] J. Kim, G. Gu, M. Park, S. Park, and J. Choo, "StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On," *Thecvf.com*, pp. 8176–8185, 2024, Accessed: Jun. 11, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Kim_StableVITON_Learning_Semantic_Correspondence_with_Latent_Diffusion_Model_for_Virtual_CVPR_2024_paper.html.

[7] X. Gu *et al.*, "Recurrent Appearance Flow for Occlusion-Free Virtual Try-On," *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 20, no. 8, pp. 1–17, Apr. 2024, doi: https://doi.org/10.1145/3659581.

[8] https://huggingface.co/spaces/Kwai-Kolors/Kolors-Virtual-Try-On (accessed Jan. 17, 2025).

[9] https://vtry.io/ (accessed Feb. 10, 2025).

[10] https://flux1.ai/virtual-try-on (accessed Feb. 27, 2025).

[11] GoGoDuck912, "GitHub - GoGoDuck912/Self-Correction-Human-Parsing: An out-of-box human parsing representation extractor.," *GitHub*, 2019. https://github.com/GoGoDuck912/Self-Correction-Human-Parsing (accessed Feb. 15, 2025).

[12] Hzzone, "GitHub - Hzzone/pytorch-openpose: pytorch implementation of openpose including Hand and Body Pose Estimation.," *GitHub*, 2018. https://github.com/Hzzone/pytorch-openpose.git (accessed Feb. 21, 2025).

[13] https://pytorch.org/ (accessed March 21, 2025).

[14] "Dress Code Dataset," *GitHub*, Feb. 05, 2023. https://github.com/aimagelab/dress-code.

[15] "GitHub - albumentations-team/albumentations: Fast image augmentation library and an easy-to-use wrapper around other libraries. Documentation: https://albumentations.ai/docs/ Paper about the library: https://www.mdpi.com/2078-2489/11/2/125," *GitHub*. https://github.com/albumentations-team/albumentations.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Apr. 2022. Available: https://arxiv.org/pdf/2112.10752.

[17] D. Kingma and M. Welling, "Auto-Encoding Variational Bayes," 2014. Available: https://arxiv.org/pdf/1312.6114.

[18] R. Gal *et al.*, "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion." Available: https://arxiv.org/pdf/2208.01618.

[19] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," 2021. Available: https://arxiv.org/pdf/2103.00020.