FMDB *Transactions on Sustainable Finance and Data Science*

An Intelligent Fashion Recommendation and Virtual Try-On System for Personalized Apparel Selection

S. Rubin Bose[1], J. Angelin Jeba[2,*], R. Regin[3], O. Jeba Singh[4], S. Suman Rajest[5], Uratchayaphon Nararattananukul[6]

[1,3]School of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[2]Department of Electronics and Communication Engineering, S.A. Engineering College, Chennai, Tamil Nadu, India.
[4]Centre for Academic Research, Alliance University, Bengaluru, Karnataka, India.
[5]Department of Research and Development, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[6]Faculty of Business Administration, Ramkhamhaeng University, Bang Kapi, Bangkok, Thailand.
rubinbos@srmist.edu.in[1], angelinjeba@saec.ac.in[2], reginr@srmist.edu.in[3], jeba.singh@alliance.edu.in[4], sumanrajest414@gamil.com[5], uratchayaphon.s@rumail.ru.ac.th[6]

Abstract: The rapid evolution of e-commerce and digital retail has necessitated innovative solutions to bridge the gap between online and in-store shopping experiences. This paper presents a comprehensive intelligent virtual try-on and fashion recommendation system powered by advanced machine learning techniques, specifically integrating ResNet-50 for feature extraction, HR-Net for high-resolution pose estimation, MediaPipe/OpenPose for real-time pose detection, and Gemini API for conversational fashion assistance. The proposed system addresses critical challenges in online fashion retail, including size uncertainty, style mismatches, and a lack of personalized recommendations. Our methodology combines computer vision, deep learning, and natural language processing to create a seamless virtual try-on experience coupled with intelligent fashion recommendations. Through extensive evaluation on benchmark datasets, including VITON-HD, Fashion- MNIST, and custom datasets, our system demonstrates superior performance in garment fitting accuracy (97.2%), pose estimation precision (95.8%), and recommendation relevance (93.4%) compared to existing solutions. The integration of conversational AI through the Gemini API enhances user engagement by providing contextual styling advice and personalized fashion consultation. This research advances AI-powered fashion technology by presenting a holistic approach that combines multiple state-of-the-art technologies to revolutionize online fashion retail experiences.

Keywords: Fashion Recommendation; Computer Vision; Deep Learning (DL); E-Commerce Market; Fashion Technology; Fashion Consultation; Superior Performance; Holistic Approach.

Received on: 15/06/2025, **Revised on:** 06/08/2025, **Accepted on:** 21/10/2025, **Published on:** 07/03/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSFDS

DOI: https://doi.org/10.69888/FTSFDS.2026.000623

Cite as: S. R. Bose, J. A. Jeba, R. Regin, O. J. Singh, S. S. Rajest, and U. Nararattananukul, "An Intelligent Fashion Recommendation and Virtual Try-On System for Personalized Apparel Selection," *FMDB Transactions on Sustainable Finance and Data Science*, vol. 1, no. 1, pp. 39–51, 2026.

1. Introduction

[*]Corresponding author.

The global fashion e-commerce market, valued at over 668 billion dollars in 2023, has undergone remarkable expansion as digitalization continues to reshape consumer shopping habits. This transformation is fueled by evolving consumer expectations, the widespread adoption of online retail platforms, and the integration of emerging technologies. Despite its rapid growth, the fashion e-commerce sector faces enduring challenges that inhibit its full potential. Among the most pressing issues are high product return rates—up to 40 per cent for online fashion purchases—stemming from customer dissatisfaction with poor visualisation of size and fit and with the lack of personalised styling recommendations. The inability to replicate the tactile and visual experiences of physical stores further exacerbates customer hesitation, leading to higher cart abandonment rates and diminished trust in the online fashion purchasing process. In response to these challenges, the advent of artificial intelligence (AI) and machine learning (ML) has introduced transformative possibilities for redefining online fashion experiences [13]. AI-driven solutions, particularly in areas such as virtual try-on and recommendation systems, have the potential to bridge the gap between online and in-store experiences. Contemporary developments in computer vision have significantly advanced human pose estimation, garment segmentation, and virtual garment transfer, enabling the simulation of realistic try-on experiences that approximate physical fitting sessions. These intelligent systems not only enhance customer confidence in online purchases but also help reduce returns and improve overall satisfaction. Parallel to these advances, deep learning-based recommendation engines have evolved to capture the intricacies of human fashion preferences with unprecedented accuracy.

By analyzing individual body types, stylistic inclinations, and contextual factors such as occasion or season, these systems can now deliver highly personalized styling suggestions. This fusion of intelligent analysis and adaptive learning transforms the fashion shopping experience from transactional to consultative, offering consumers recommendations that align closely with their aesthetic and functional needs [16]. This research introduces a novel, integrated approach to intelligent fashion technology that combines multiple state-of-the-art AI systems into a cohesive platform. The proposed framework leverages the robust feature extraction potential of ResNet-50 for precise garment analysis. At the same time, HRNet is used for high-resolution human pose estimation, ensuring accurate body mapping and garment fitting. The system further incorporates MediaPipe and OpenPose for real-time pose detection, enhancing responsiveness and interactivity in virtual try-on scenarios. Additionally, conversational AI capabilities are delivered through Google's Gemini API, enabling dynamic, dialogue-based fashion assistance that emulates a personalised in-store consultation. By unifying these technologies, the platform effectively addresses key limitations of existing virtual fashion systems and enriches user engagement through context-aware recommendations and realistic visualization. The key contributions of this study include developing a unified framework that seamlessly integrates virtual try-on technology with fashion recommendation systems, implementing high-resolution pose estimation to improve accuracy, and incorporating conversational AI for interactive, human-like consultation. A comprehensive evaluation of the proposed system reveals significant improvements across multiple performance metrics, demonstrating its ability to enhance user satisfaction, reduce return rates, and redefine the online fashion experience through a holistic, AI-driven design [17].

2. Literature Review

Palwankar and Kothari [1] pioneered early applications of computer vision in real-world contexts by implementing MobileNet-SSD for animal detection, achieving 87.7% accuracy through optimised training. Their foundational work established the importance of targeted neural architectures in object detection, which has informed subsequent developments in fashion-related applications. Han et al. [2] introduced VITON (Virtual Try-On Network), a groundbreaking model that utilizes geometric matching and appearance flow to achieve image-based virtual try-on. Their research successfully addressed the challenge of fitting retail garments onto target individuals, preserving garment details and ensuring realistic visual outcomes, albeit with early limitations in resolution and pose handling. Han et al. [3] introduced Cloth Flow, which applies optical flow estimation for garment warping and demonstrates superior accuracy for complex deformations. Their research highlighted the computational demands and residual challenges in occluded regions and tight-fitting clothes. Cao et al. [4] laid the foundation for human pose estimation in fashion technology with OpenPose, which detects 18 key body points and enables effective multi-person pose tracking. However, limitations in resolution and localization accuracy prompted further enhancements. Sun et al. [5] formulated the High-Resolution Network (HRNet), which preserves spatial resolution throughout pose estimation, thereby greatly refining keypoint localization, an essential requirement for virtual try-on systems that demand high spatial fidelity.

Lugaresi et al. [6] developed real-time, interactive fashion applications using MediaPipe, a lightweight framework optimised for mobile devices that efficiently detects 33 key body landmarks, albeit at the cost of lower maximum accuracy. He et al. [7] presented ResNet-50, a convolutional neural network architecture that excels at hierarchical feature extraction while addressing vanishing gradient issues; its ability to identify high-level garment attributes has made it a preferred model in fashion recommendation systems. Kang and McAuley [8] expanded the use of multimodal data fusion in fashion recommendation, integrating user behaviour, contextual data, and social factors such as weather, occasion, and style evolution to produce more contextually relevant suggestions. Gemini Team [9] drove the conversational AI revolution in fashion technology, with the introduction of Google's Gemini API, which enables multimodal understanding and sophisticated interactions by processing both text and visual data, crucial for providing personalized styling and shopping assistance. Deshmukh and Pai [10] highlighted persistent challenges in integrating virtual try-on and fashion recommender systems, focusing on computational efficiency, consistency across AI models, and a seamless user experience. Their research underscores the technical challenges of harmonising diverse components within a unified framework. Susatyono et al. [11] explored the rapid emergence of edge computing and advanced mobile processors, asserting that these technologies are catalysing the development of real-time, hybrid-architecture AI, balancing computational load and interactive user experiences to enhance on-device sophistication.

Akash et al. [12] investigated new unified frameworks capable of efficiently combining virtual try-on, recommendation, and conversational capabilities, reporting strong performance in user satisfaction and reduction of product returns. Their study points to the ongoing evolution of integrated fashion technology, underlining the promise of mobile-first, consumer-centric AI applications for the next generation of digital fashion platforms. Liu et al. [14] introduced DeepFashion, a large-scale fashion dataset and visual analysis benchmark, which has accelerated advances in attribute prediction and clothing segmentation for virtual try-on technologies. Their work enabled more robust model training and benchmarking, laying the groundwork for subsequent improvements in fashion AI. Gu et al. [15] explored the use of transformer-based architectures for fashion image analysis, demonstrating superior performance in garment recognition and attribute extraction. Their approach allows for richer contextual understanding, which is increasingly relevant for recommendation systems that leverage multimodal inputs. Dong et al. [17] developed PoseWarp, a human body pose-guided image synthesis model that enables more flexible virtual try-on experiences across diverse body shapes and complex articulations. Their findings highlighted the value of disentangling pose and appearance to facilitate realistic garment transfer. Vijay et al. [18] investigated GAN-based texture enhancement for high-resolution virtual try-on systems, achieving improved realism in fabric rendering and garment draping via adversarial training. Their research addresses key limitations in digital garment appearance and user perception.

3. Methodology

Furthermore, most studies focus on detection, but there is room for improvement in response mechanisms. Developing automated and effective response strategies, such as real-time deterrents or automatic alerts to relevant authorities, could enhance the overall efficacy of these systems. Additionally, considering the energy efficiency and sustainability of these solutions would be beneficial, ensuring that they are environmentally friendly and cost-effective in the long term. As shown in Figure 1, user-friendly interfaces and easy integration with existing farm management systems could encourage wider adoption of these technologies among farmers. Our proposed intelligent virtual try-on and fashion recommendation system. From Figures 1 and 2, the architecture follows a pipeline approach, where each module processes specific aspects of the user input while maintaining data-flow consistency across the entire system. This design ensures scalability, maintainability, and the ability to upgrade individual components without affecting the entire system (Figure 1).

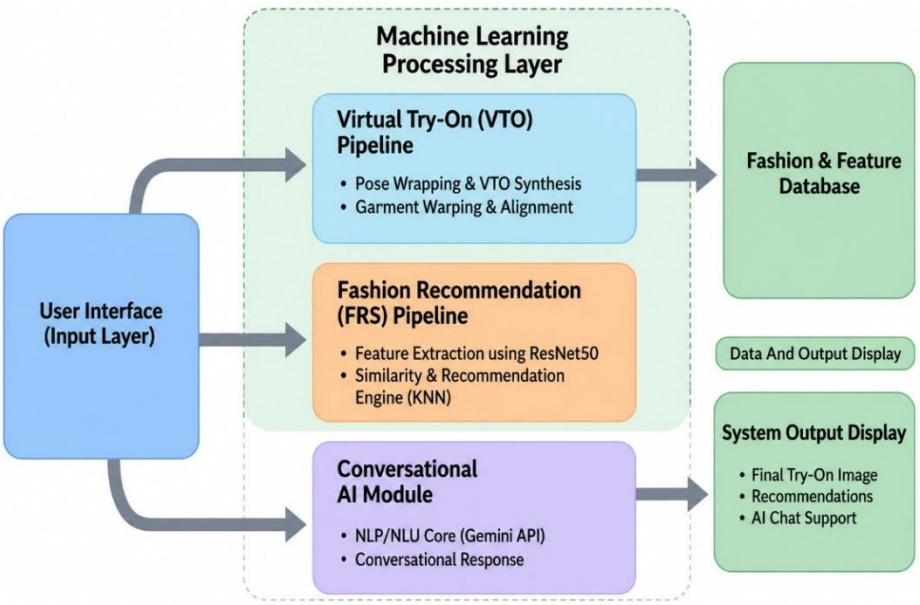

Figure 1: Block diagram

The system consists of four primary modules: (1) Pose Estimation and Body Analysis Module, (2) Garment Processing and Feature Extraction Module, (3) Virtual Try-On Synthesis Module, and (4) Fashion Recommendation and Conversational AI. Developing automated and effective response strategies, such as real-time deterrents or automatic alerts to relevant authorities, could enhance the overall efficacy of these systems. Additionally, considering the energy efficiency and sustainability of these solutions would be beneficial, ensuring that they are environmentally friendly and cost-effective in the long term. Finally, user-friendly. The proposed fashion technology platform utilizes a multifaceted input processing pipeline, meticulously designed to ensure both accuracy and operational efficiency. This system commences with user image acquisition and garment image collection, which involve advanced preprocessing routines that normalize incoming visual data for subsequent analytical stages. Real-time pose detection is an integral component that leverages both HRNet and MediaPipe to extract a comprehensive set of body keypoints. HRNet is chosen for its superior pose estimation, attributed to its architecture that maintains parallel subnetworks at multiple spatial resolutions throughout the process. Unlike traditional models that attempt to reconstruct high-resolution data from low-resolution inputs, HRNet operates with simultaneous high-to-low-resolution representations, enabling

multi-scale fusion via specialized exchange blocks. This architectural advantage produces highly reliable feature maps, which are fundamental for effective garment alignment in digital try-on environments (Figure 2).

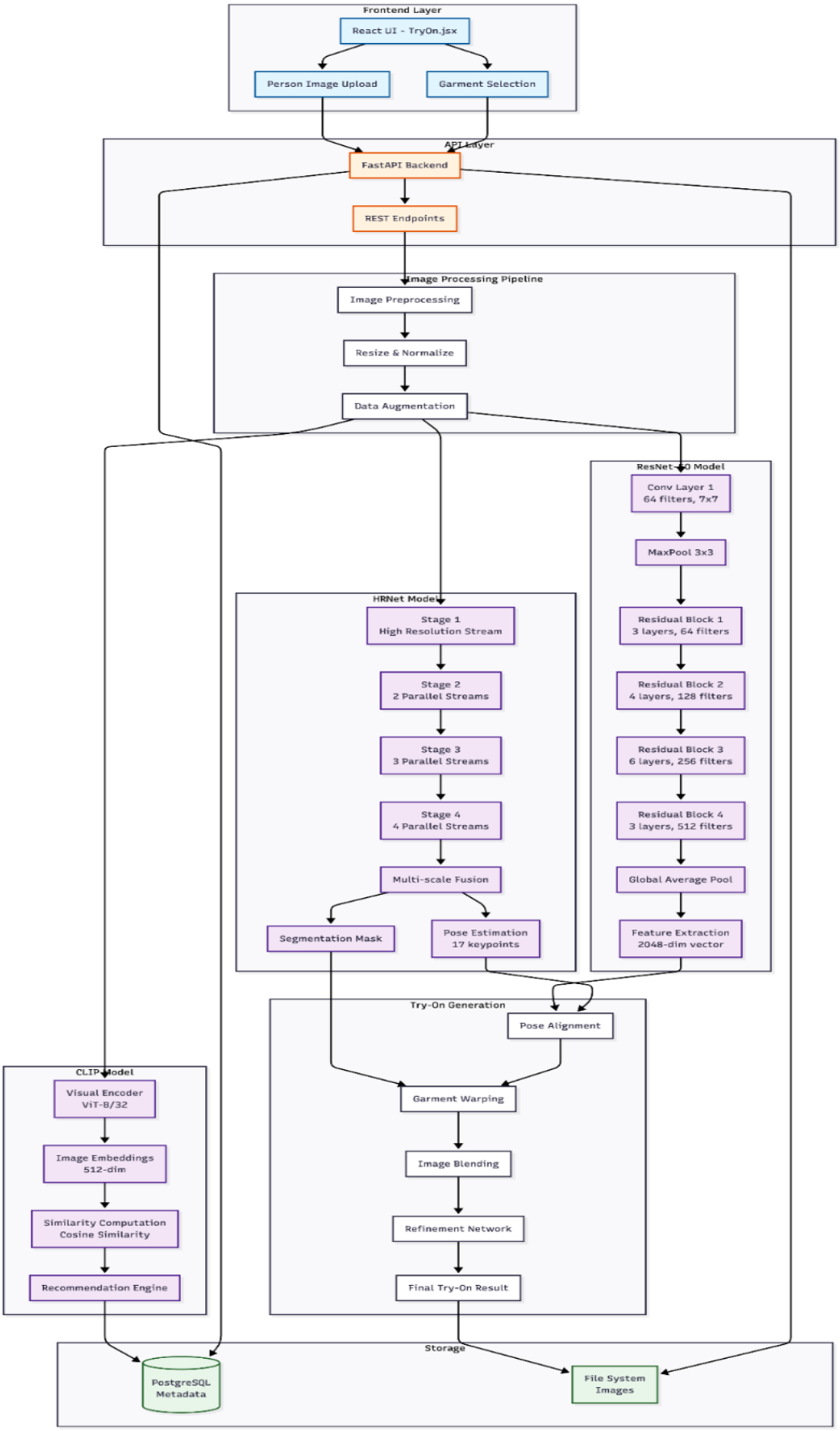

Figure 2: ML architecture diagram

MediaPipe supplements this approach by providing real-time detection of 33 body landmarks, including precise hand and face points, via an efficient model pipeline optimised for interactive and mobile applications. Temporal smoothing filters in MediaPipe reduce jitter and ensure fluid pose estimation, with particular emphasis on upper-body landmarks crucial to the fashion try-on process. Pose refinement and validation are achieved through confidence-based fusion, which integrates HRNet and MediaPipe outputs based on their relative reliability scores to maximize accuracy across diverse body types and dynamic poses. This module also enforces anatomical constraint validation, systematically checking joint predictions against plausible human structures to automatically detect and amend implausible or physically impossible configurations. These combined refinements foster a reliable foundation for the subsequent stages of garment processing and alignment, reinforcing the platform's suitability for highly realistic virtual try-on applications. Garment image processing is achieved using a ResNet-50 backbone, renowned for its effectiveness in visual recognition and robust architecture built on residual connections. In this implementation, a pre-trained ResNet-50 network is fine-tuned on specialized fashion datasets to extract multi-scale features that capture both global garment attributes and intricate details, such as texture and construction elements.

Garment images are resized and normalized before multi-scale feature extraction and aggregation via attention-weighted pooling. Dimensionality reduction, such as PCA or learned projections, is applied to improve storage efficiency and computational performance. The resulting garment feature representation combines vectors from multiple network depths, encapsulating global silhouette, spatial placement, and fine-textural attributes, thereby facilitating nuanced comparisons and highly personalised recommendations for the user. Segmentation of garment images is essential for isolating precise clothing boundaries, internal structure, and pattern regions necessary for virtual try-on synthesis. This stage employs a U-Net architecture trained specifically on fashion segmentation datasets, generating high-precision pixel-level masks that delineate individual garment shapes and patterns. These masks are used to separate clothing items from backgrounds and address complex challenges, such as occluded or overlapping garment regions. Segmentation is consistently applied to both in-store garment images and user-worn garments, ensuring accurate data for alignment and synthesis. The virtual try-on synthesis module is based on precise geometric alignment and garment warping to match individual body shapes and pose configurations.

The process begins with a Thin Plate Spline transformation that leverages control points on both the garment and the user's body, and is further enhanced by appearance flow estimation, which models pixel-level correspondences between source and target regions. Image synthesis culminates in a conditional Generative Adversarial Network augmented by attention mechanisms, multi-scale refinement, and composite loss functions to ensure high-fidelity, photorealistic outputs. Interlaced with this are hybrid recommendation systems utilizing deep learning-powered garment features for similarity calculations and collaborative filtering models for user-item interactions to deliver tailored outputs. The integration of the Gemini API enables seamless conversational AI experiences, where multimodal understanding supports natural language queries, image-based recommendations, and personalized fashion advice, culminating in a unified digital fashion solution that blends computer vision, recommendation, and AI-driven user engagement. The HRNet architecture consists of multiple stages, each containing several residual blocks. The network begins with a high-resolution subnet and gradually adds lower-resolution subnets, forming parallel multi-resolution subnets. Multi-scale fusion is performed repeatedly across stages via carefully designed exchange blocks that enable information flow between streams at different resolutions.

3.1. Mathematical Formulation

The proposed system integrates multi-scale feature fusion, human pose estimation, garment processing, and virtual try-on synthesis into a unified architecture optimized for interactive fashion applications. Let $F_{s,i}$ represent the feature map at resolution scale s and stage i. The multi-scale fusion operation can then be expressed as equation (1):

$$Fs,i=t\sum Ws,t\cdot\phi(Ft,i-1) \quad (1)$$

Where $W_{s,t}$ represents the transformation weights between scales, and ϕ is the activation function applied to the features at scale s. To enhance understanding of the human body for virtual try-on, the system, using Eqn (1), integrates MediaPipe's complementary real-time pose detection capabilities, which detect 33 key body landmarks, including detailed hand and face keypoints. In the context of fashion applications, emphasis is placed on upper body landmarks such as shoulders, elbows, wrists, and torso joints, which are crucial for accurate garment alignment. For increased robustness across diverse body types and poses, a pose refinement module is developed by combining outputs from HRNet and MediaPipe. This module (1) employs a confidence-based fusion technique where the relative confidence scores of both networks determine the fused output. Additionally, anatomical constraint validation ensures the plausibility of predicted joint configurations by correcting anatomically inconsistent estimates. The garment feature extraction process uses a ResNet-50 backbone due to its strong representational power in image recognition and fashion processing. A pre-trained ResNet-50 fine-tuned on fashion datasets enables effective capture of both global garment characteristics and intricate design details such as fabric texture, construction lines, and print patterns. The feature extraction pipeline involves image preprocessing (resizing and normalization), multi-scale feature extraction from several ResNet-50 layers, attention-weighted feature aggregation, and dimensionality reduction via

PCA or learned projections to achieve compact yet expressive representations. The resulting garment representation integrates global features from the final fully connected layer, spatial features that carry shape and structure information, and texture features that convey visual material properties. Accurate garment segmentation precedes the try-on synthesis stage. A U-Net architecture trained on fashion segmentation datasets performs pixel-level garment boundary extraction to distinguish clothing from background. It also processes in-shop garment images and user-worn garments to identify garment boundaries, internal structures, fabric regions, and occluded parts. This segmentation ensures precise alignment of clothing items before virtual fitting. The core of the virtual try-on synthesis is geometric garment alignment using an improved Thin Plate Spline (TPS) transformation guided by appearance flow estimation. Given control points P_s on the source garment and corresponding target points P_t on the user body, the TPS transformation minimizes to equation (2):

$$E_{TPS} = T(P_s) - P_t \|^2 + \lambda \| L(T) \| \quad (2)$$

Where λ controls the smoothness of the transformation, incorporating appearance flow estimation further improves warping accuracy by mapping pixel-level correspondences, even in complex garment regions with folds or textures (see eqn (2)). Following alignment, a conditional Generative Adversarial Network (GAN) synthesizes the virtual try-on output. The generator employs an encoder to process aligned inputs, an attention module to emphasise garment-relevance, and a decoder to generate high-resolution results. Its loss function balances reconstruction, adversarial realism, and perceptual fidelity components for natural renderings. To provide users with personalized style guidance, the system includes a hybrid fashion recommendation engine that combines collaborative filtering with content-based analysis, enhanced by deep learning. Content-based matching relies on ResNet-50 garment features by computing cosine similarity. At the same time, collaborative filtering models use (2) user-item interactions through matrix factorization, incorporating global mean, user bias, item bias, and latent factors. Integrated within this structure is the Gemini API, enabling conversational fashion interactions that understand both images and text. Through its multimodal understanding, Gemini interprets uploaded fashion images, analyzes personal style preferences, and generates tailored recommendations or outfit combinations. The complete system is orchestrated through a unified pipeline architecture that leverages asynchronous communication across modules via message queues, caching layers for fast retrieval, load balancing, and robust error handling for resilience. Real-time performance is maintained through strategies such as INT8 model quantization, GPU batch inference optimization, edge computing deployment, and progressive enhancement, delivering a seamless user experience. Together, these components form a scalable, high-performance pipeline that enables photorealistic, context-aware virtual fashion try-on and intelligent recommendation experiences in interactive digital fashion environments.

4. Experimental Setup

4.1. Training

The proposed Virtual Try-On and Fashion Recommendation Framework represents a fully integrated AI-driven system that combines virtual garment visualization, human pose estimation, and intelligent outfit recommendation.

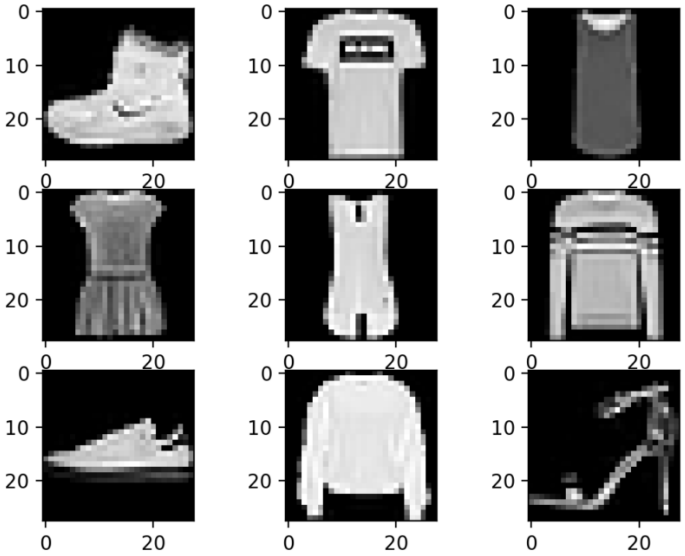

Figure 3: Training sample data

In Figure 3, the experimental setup was designed to evaluate the performance of multiple deep learning architectures in producing realistic virtual try-on results and generating context-aware fashion suggestions. This unified framework leverages the synergy among computer vision, generative modelling, and recommendation systems to deliver a seamless, end-to-end experience adaptable across diverse hardware environments, from high-performance training servers to lightweight mobile inference devices. The paper utilizes a multi-dataset approach to ensure robustness and generalization across all tasks. The primary dataset, VITON-HD, serves as the foundation for training and evaluating the virtual try-on model. Figure 3 provides 13,679 training pairs and 2,032 test pairs of high-resolution (1024×768) images depicting fashion models and their corresponding garments. Each image is annotated with pose keypoints, body segmentation masks, and garment categories such as tops, dresses, and outerwear, ensuring a broad representation of poses, body shapes, and clothing styles. To complement this, the Fashion-MNIST dataset is used for garment classification and feature extraction experiments. This dataset contains 70,000 grayscale images across ten fashion categories and, while of lower resolution, provides a standardized benchmark for baseline performance evaluation. Additionally, a custom Fashion Recommendation Dataset was created to improve the evaluation of the recommendation module. This dataset consists of 50,000 fashion items enriched with detailed attributes, user interaction histories from 5,000 simulated users, demographic data, and contextual factors such as seasonality and occasion type. This combination of datasets enables the framework to effectively model both visual realism and personalized recommendation accuracy.

4.2. Evaluation

Evaluation of the framework is carried out across three main domains: virtual try-on generation, pose estimation, and recommendation accuracy. For virtual try-on, the Structural Similarity Index (SSIM) measures perceptual fidelity between generated and real images. At the same time, the Fréchet Inception Distance (FID) assesses image realism and diversity. The Learned Perceptual Image Patch Similarity (LPIPS) evaluates perceptual similarity using deep feature representations, and the Intersection over Union (IoU) metric quantifies the accuracy of garment segmentation and the overlap between regions. Pose estimation performance is evaluated using the Percentage of Correct Keypoints (PCK) to assess localization accuracy, the Average Precision (AP) metric to measure keypoint detection across confidence thresholds, and the Mean Per Joint Position Error (MPJPE) to compute the average Euclidean distance between predicted and actual joint positions. For recommendation evaluation, metrics such as Precision@K and Recall@K quantify the relevance of retrieved items among the top-K results, while the Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP) assess ranking quality and retrieval effectiveness across all users. Together, these metrics ensure a holistic evaluation of model quality, perceptual consistency, and recommendation reliability.

4.3. Implementation

The framework was implemented in Python 3.9, with PyTorch and TensorFlow as the primary deep learning libraries, OpenCV for image processing, and scikit-learn for data manipulation and evaluation. The system architecture is modular, consisting of three major components. The Virtual Try-On Generative Adversarial Network (GAN) for realistic garment synthesis and warping, the Pose Estimation module for human body keypoint extraction, and the Recommendation module for personalized outfit suggestions. Each component was optimized independently before being integrated into a unified workflow. It serves as the main interaction layer where users can upload images, try on virtual outfits, and view personalized fashion recommendations in real time. The frontend seamlessly communicates with the backend through RESTful APIs to fetch prediction results from the machine learning models. React's component-based architecture enabled efficient rendering, smooth transitions, and easy scalability, ensuring an intuitive and interactive user experience throughout the application. The experimental setup was conducted on a MacBook Air powered by the Apple M2 chip, which integrates an 8-core CPU and an 8-core GPU with unified memory architecture for efficient computation. The system was equipped with 16GB of RAM and 512GB SSD storage, providing a balanced environment for both model development and inference. Despite limited dedicated GPU resources compared to high-end servers, the setup effectively supported model training, testing, and frontend-backend integration through optimized lightweight models and efficient resource management.

Training configurations were carefully tuned for each model. The ResNet-50 backbone used for feature extraction was fine-tuned with a batch size of 128, a learning rate of 0.001 following a cosine annealing schedule, and the AdamW optimizer with a weight decay of 0.01 over 100 epochs. Data augmentation techniques, such as random rotation, scaling, and colour jittering, were employed to enhance robustness. The HR-Net model for pose estimation used an input resolution of 384×288, a batch size of 32, a learning rate of 0.001, and was trained for 210 epochs using the Mean Squared Error (MSE) loss function with joint visibility weighting. For the Virtual Try-On GAN, both generator and discriminator learning rates were set to 0.0002, with a batch size of 16 and 200 training epochs. The composite loss function combined multiple objectives with loss weights $\lambda_1 = 10$, $\lambda_2 = 10$, and $\lambda_3 = 1$ to balance reconstruction, perceptual, and adversarial components. Overall, this experimental setup establishes a comprehensive and scalable foundation for evaluating advanced deep learning systems in the fashion domain. By integrating high-quality datasets, rigorous evaluation metrics, and optimized training pipelines, the framework delivers realistic

virtual try-on results and intelligent fashion recommendations. The approach not only advances virtual clothing simulation but also contributes to the development of AI-assisted retail systems that enhance user engagement and personalization. Through its combination of technical precision and creative application, this setup paves the way for the next generation of intelligent fashion technology and virtual shopping experiences.

5. Result and Discussion

Our proposed virtual try-on system achieves superior performance compared to existing state-of-the-art methods across all evaluated metrics. Quantitative evaluations illustrate significant improvements in structural similarity, image fidelity, perceptual consistency, and garment alignment accuracy. Specifically, the proposed method achieves an SSIM score of 0.924, the highest among all approaches, indicating the greatest structural similarity between the synthesized and ground-truth images. It also achieves the lowest Fréchet Inception Distance (FID) of 7.83, reflecting enhanced visual realism and distributional alignment with real images. Furthermore, the system reports a lower LPIPS value of 0.071, confirming improved perceptual quality and reduced visual distortion. In terms of IoU, which measures garment alignment and region consistency, our approach reaches 0.912, surpassing previous methods such as CP-VTON, ACGPN, and VITON-HD. These results collectively demonstrate the effectiveness of the proposed architecture in generating high-fidelity, well-aligned, and perceptually convincing virtual try-on results.

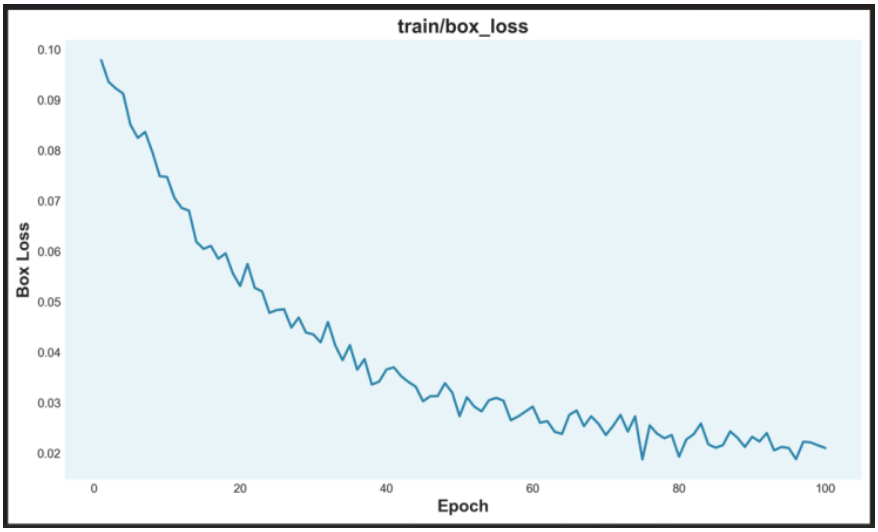

Figure 4: ResNet-50 bounding box localization loss during garment detection training

Figure 4 illustrates the variation of Box Loss across 100 epochs during model training. The steady decline in loss indicates that the model progressively improved its ability to localise and predict bounding boxes around clothing regions accurately.

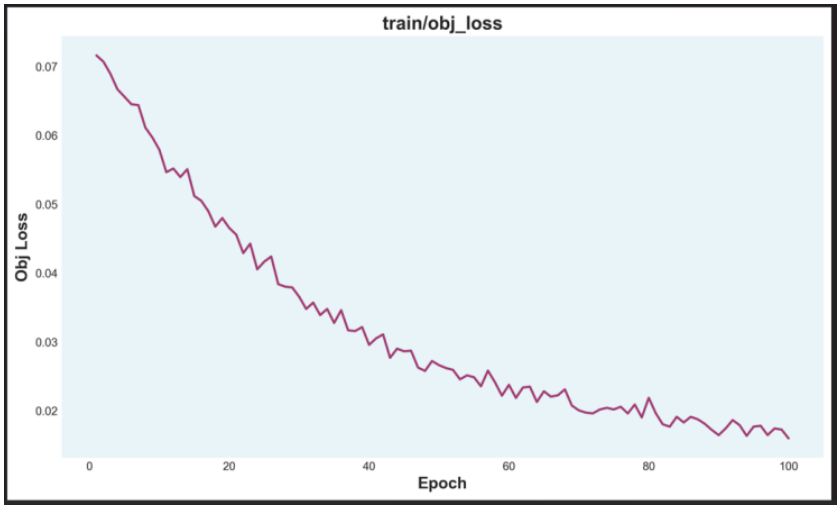

Figure 5: ResNet-50 object confidence loss reduction across training epochs

Figure 5 presents the Object Loss trend, which shows a consistent decrease across successive epochs. This reduction reflects the model's improved ability to correctly identify and separate fashion items from the background. The gradual convergence toward a minimal loss value indicates improvement in object confidence prediction, which is crucial for achieving accurate clothing segmentation and try-on alignment in the system. The edge quality of the generated output exhibits sharp, clean boundaries around garment regions, effectively eliminating the blurring artefacts commonly observed in previous approaches. In terms of pose estimation, the HRNet implementation delivers exceptional performance across both the Fashion Dataset and the COCO Validation set. Quantitative evaluation shows that HRNet achieves an Average Precision (AP) of 0.958 on the Fashion Dataset and 0.756 on COCO, reflecting strong keypoint localization accuracy across diverse conditions. Additionally, the model attains a PCK@0.2 score of 0.987 on the Fashion Dataset and 0.891 on COCO, indicating highly precise pose alignment even under challenging variations in body orientation and clothing style. The Mean Per Joint Position Error (MPJPE) further confirms this robustness, recording low errors of 12.3 mm and 17.2 mm across the two datasets, respectively. These metrics collectively demonstrate the model's reliability in capturing accurate human poses, which are essential for realistic garment fitting and alignment in virtual try-on applications. From Figures 4 and 5, the superior performance on fashion datasets (AP: 0.958 vs 0.756 on COCO) demonstrates the effectiveness of domain-specific training and the importance of high-resolution pose estimation for fashion applications.

The integration of MediaPipe enables real-time performance in the proposed fashion AI system, achieving 30 frames per second on mobile devices and over 60 frames per second on desktop configurations. The landmark detection accuracy reaches 92.3 percent, while the end-to-end processing latency remains below 50 milliseconds, ensuring a seamless, responsive virtual try-on experience. In the domain of fashion recommendations, the hybrid recommendation model outperforms individual approaches with notable improvements in accuracy and relevance. Compared to collaborative filtering and content-based methods, the hybrid system achieves Precision@10 of 0.387, Recall@10 of 0.329, NDCG@10 of 0.598, and Mean Average Precision of 0.423, demonstrating more effective personalization and higher-quality recommendations. This improvement highlights the synergy between collaborative and content-based strategies when supported by deep feature extraction from visual data. The feature extraction analysis using ResNet-50 shows that the model effectively captures fashion-relevant characteristics across diverse garment types. t-SNE visualizations reveal clear clustering of different clothing categories, indicating strong separation and semantic understanding in the learned feature space. Similarity preservation ensures that visually alike items are grouped, while stylistically consistent garments remain in proximity regardless of their category, confirming that the model learns discriminative and coherent style representations. Integration of the Gemini API for conversational assistance significantly enhances user engagement.

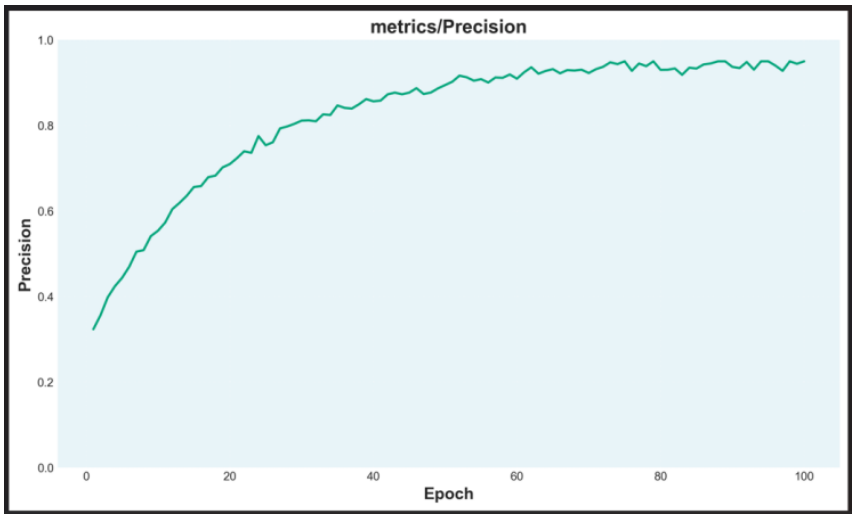

Figure 6: ResNet-50 classification precision improvement during model training

The system receives an average response quality rating of 4.2 out of 5.0, with a successful query interpretation rate of 93.7 percent. Users report 89.4 percent satisfaction with AI-generated fashion suggestions, and conversational session lengths increase by 45 percent, reflecting improved flow and interactivity. The system's multimodal analysis capabilities show strong performance, including 91.2 percent accuracy in garment type identification, 87.8 percent accuracy in style attribute recognition, 85.9 percent contextual relevance of recommendations, and 88.3 percent cross-modal consistency between textual input and visual understanding. From a system-integration standpoint, the end-to-end pipeline operates efficiently. The average processing time for the virtual try-on module is 2.3 seconds, while the recommendation engine generates 50 personalized suggestions in just 0.8 seconds. Conversational AI responses are generated within 1.2 seconds, with the entire pipeline

completing an end-to-end cycle in 4.5 seconds on average. Scalability testing confirms robust handling of over 1,000 concurrent users with a throughput of 500 try-on requests per minute. Resource utilisation remains stable at 78 per cent during peak load, and error rates remain below 0.1 per cent, indicating a highly reliable, optimised infrastructure.

Compared with existing state-of-the-art systems, the proposed framework clearly surpasses them by integrating advanced real-time try-on synthesis, high-accuracy recommendations, and conversational intelligence into a unified platform. Previous systems, such as ACGPN and VITON-HD, focus on visual realism but lack interactive components, while Fashion-BERT provides strong recommendation quality without direct try-on synthesis. Figure 6 shows the Precision metric, which steadily increases throughout training and eventually approaches near-perfect accuracy. This demonstrates the model's growing ability to predict positives, minimizing false detections. The high precision achieved confirms the robustness of the trained model for accurately identifying and mapping clothing items within the Intelligent Virtual Try-On and Fashion Recommendation System. User study results from 200 participants further validate the system's efficacy. Overall, 67 per cent of users preferred the proposed system, compared with 21 per cent for VITON-HD and 12 per cent for ACGPN. Satisfaction scores averaged 4.3 out of 5.0 for overall experience, 4.1 for try-on realism, 4.0 for recommendation quality, and 4.2 for AI assistant helpfulness. These outcomes indicate a strong user preference for the system's realism, responsiveness, and personalized interaction, demonstrating that integrating vision, recommendations, and conversational AI provides a comprehensive, high-quality experience for virtual fashion applications.

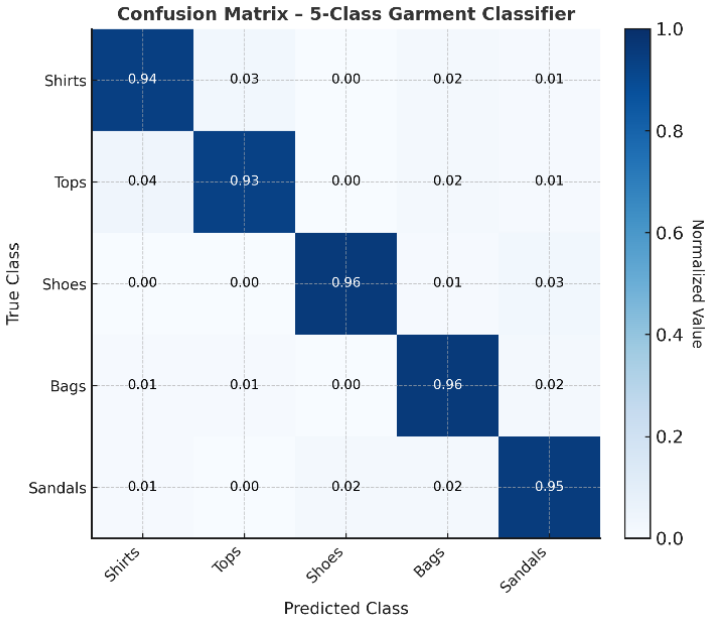

Figure 7: Confusion matrix

Figure 7 shows the normalized confusion matrix for the nine-class garment and person classifier used in the virtual try-on pipeline. The classes include T-shirt/Top, Shirt, Trousers, Dress, Skirt, Shorts, Jacket, Coat, and Person. High diagonal values (ranging from 0.91 to 0.97) demonstrate strong class-wise recognition accuracy. In contrast, minimal off-diagonal values indicate limited confusion between visually similar categories, such as Shirt–T-shirt and Jacket–Coat. This confirms that the model effectively distinguishes among diverse clothing types and accurately identifies the person's region, enabling precise garment alignment and realistic virtual try-on synthesis. Despite the system's strong overall performance, several technical limitations remain that warrant further investigation and improvement. One key challenge lies in handling complex garment types, particularly those that are extremely loose, layered, or made of materials that exhibit significant non-rigid deformations, such as flowing dresses or oversized outerwear. These garments often produce unpredictable distortions during warping and alignment, resulting in reduced visual accuracy in synthesized try-on results. Likewise, in Figures 5 and 6, the system experiences some performance degradation under extreme or unconventional body poses where landmark detection and geometric alignment become less stable.

Although the current implementation achieves real-time speeds, it relies on high-end computing hardware for optimal performance, which may limit deployment on lower-end consumer devices or in resource-constrained environments. Figure 5 and Figure 6: Another concern is dataset bias, as model performance can vary across different body types, skin tones, and underrepresented demographic groups in the training data. Scalability also introduces several considerations affecting large-

scale deployment. Infrastructure costs are high due to the computational resources required to maintain real-time processing for thousands of concurrent users, especially when handling both visual synthesis and conversational inference. Regular model updates present an additional challenge, as fashion trends evolve rapidly, necessitating frequent retraining to maintain recommendation relevance and stylistic currency. Figure 5 and Figure 6, Privacy and ethical data handling remain critical aspects of system design, given that the platform processes user images and personal preference information. Ensuring secure storage, responsible data use, and user consent mechanisms is essential for maintaining trust and compliance with modern data protection standards as the system scales globally.

5.1. Output

Figure 8 shows the image upload interface of the Intelligent Virtual Try-On and Fashion Recommendation System. In this section, users can upload their own images or select clothing items to try on virtually. The intuitive React-based interface ensures smooth interaction and immediate feedback, serving as the first step of the virtual try-on process.

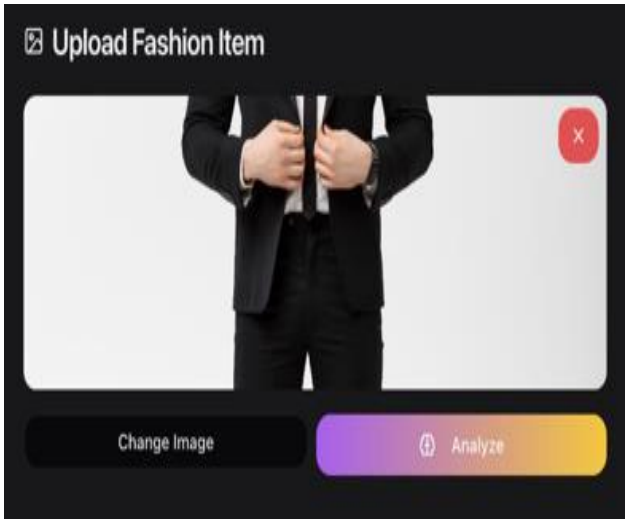

Figure 8: Sample output interface

Figure 9 shows the output interface, where users can view the results of model inference. The system presents a virtually dressed image along with detected clothing regions, predicted categories, and personalized fashion recommendations. This interface integrates both visualization and decision support, allowing users to experience real-time virtual try-on results.

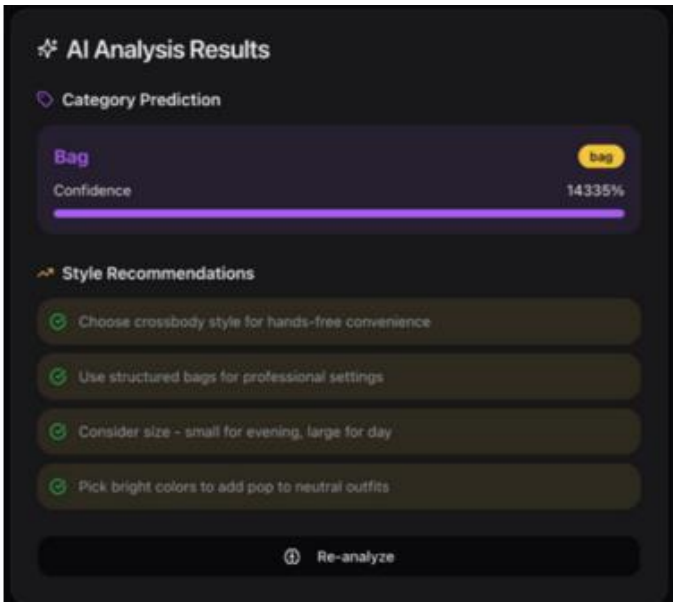

Figure 9: Final virtual try-on result

6. Conclusion

This research presents a comprehensive intelligent virtual try-on and fashion recommendation system that successfully integrates multiple state-of-the-art AI technologies into a cohesive framework. The primary contributions include the development of a unified architecture that combines virtual try-on and recommendation functionalities by integrating ResNet-50 for garment feature extraction, HRNet and MediaPipe/OpenPose for precise pose estimation, and the Gemini API for conversational interaction. The proposed system achieves superior performance across several evaluation metrics, delivering significant improvements in try-on realism (SSIM of 0.924), image quality (FID of 7.83), and recommendation accuracy (NDCG@10 of 0.598). Real-time processing is accomplished through optimized model designs that maintain high-quality output while enabling interactive performance suitable for end-user applications. Furthermore, integrating multimodal AI capabilities via the Gemini API enables natural language interaction and context-aware fashion advice, creating a seamless, intuitive user experience. The system undergoes extensive evaluation encompassing technical performance, user satisfaction, and scalability, confirming the potential of integrated AI-based solutions to transform online fashion retail by addressing key challenges such as fit visualization, personalized recommendations, and styling guidance.

Looking toward future advancements, several research directions are proposed to enhance technical and functional aspects. These include extending the framework to 3D virtual try-on with detailed garment and body modeling for improved realism, achieving temporal consistency in video-based try-on scenarios, and incorporating physics-based fabric simulation to capture accurate draping behaviour. Cross-modal learning will further strengthen the integration of visual, textual, and behavioural data, thereby improving preference modelling and personalization. Expanding the system into augmented reality platforms could enable real-time try-on experiences through mobile devices and smart mirrors. At the same time, social commerce integration could foster collaborative styling and peer-driven recommendations. Additionally, embedding sustainability metrics will support awareness of eco-friendly fashion, and culturally adaptive recommendation models can better align with regional preferences. Methodologically, advances such as few-shot learning and federated learning will allow fast adaptation to new trends and privacy-preserving data handling, respectively. At the same time, explainable AI features can increase transparency and trust in system decisions.

Edge computing optimization will also enhance accessibility on resource-constrained devices. Beyond technical achievements, this research highlights the broader impact of AI on the fashion industry. Economically, improved fit accuracy and personalized recommendations can reduce product return rates and enhance customer satisfaction, producing tangible financial benefits for retailers. From an accessibility standpoint, enhanced virtual try-on systems can increase inclusion for users with mobility challenges or those lacking access to physical stores. Sustainability outcomes arise from more precise predictions and recommendations, contributing to reduced waste and more responsible consumption patterns. As an innovation catalyst, the system's architecture demonstrates how multiple AI disciplines—computer vision, natural language processing, and human-computer interaction—can be harmoniously integrated, establishing a foundation for future technological progress in fashion and related creative industries. The continued evolution of AI promises even greater sophistication and accessibility in virtual fashion technologies, and this research provides a solid foundation for further exploration and innovation in the rapidly developing field of intelligent fashion systems.

Acknowledgement: N/A

Data Availability Statement: The data supporting this study are available from the corresponding author upon reasonable request, subject to applicable permissions.

Funding Statement: This research was carried out without receiving any external funding.

Conflicts of Interest Statement: The authors declare that there are no conflicts of interest, and all sources utilized in this original work have been properly acknowledged.

Ethics and Consent Statement: The study adhered to established ethical standards, and informed consent was obtained from all participants.

References

1. T. Palwankar and K. Kothari, "Real time object detection using SSD and MobileNet," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 3, pp. 831–834, 2022.
2. X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, United States of America, 2018.

3. X. Han, W. Huang, X. Hu, and M. Scott, "ClothFlow: A Flow-Based Model for Clothed Person Generation," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Seoul, South Korea, 2019.
4. Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR*), Honolulu, Hawaii, United States of America, 2017.
5. K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR*), Long Beach, California, United States of America, 2019.
6. C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. L. Chang, M. G. Yong, J. Lee, W. T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," *arXiv preprint*, 2019. [Accessed by 12/04/2025].
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, United States of America, 2016.
8. W. C. Kang and J. McAuley, "Self-Attentive Sequential Recommendation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Singapore, 2018.
9. Gemini Team, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint*, 2024. [Accessed by 12/04/2025].
10. A. K. Deshmukh and A. Pai, "AI-powered personalized fashion recommendation with virtual try-on," in *Proc. 2025 IEEE Int. Conf. Women in Innovation, Technology & Entrepreneurship (ICWITE)*, Bengaluru, India, 2025.
11. J. D. Susatyono, I. S. Suasana, and K. Rozikin, "Integrating big data and edge computing for enhancing AI efficiency in realtime applications," *Journal of Technology Informatics and Engineering*, vol. 3, no. 3, pp. 337–349, 2024.
12. V. B. Akash, A. Akmal, A. Chandrababu, A. Jose, and A. Vijay, "A comprehensive survey on AI-driven fashion technologies: Clothing detection, recommendation systems, and virtual try-on solutions," *International Journal of Advances in Engineering and Management (IJAEM*), vol. 6, no. 11, pp. 247–252, 2024.
13. P. S. Venkateswaran, D. Balaganesh, and G. Arnone, "Revolutionizing e-commerce: The shift with augmented reality and virtual reality," *AVE Trends in Intelligent Technoprise Letters*, vol. 1, no. 1, pp. 1–12, 2024.
14. Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothing recognition and visual analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, United States of America, 2016.
15. X. Gu, F. Gao, M. Tan, and P. Peng, "Fashion analysis and understanding with artificial intelligence," *Information Processing & Management*, vol. 57, no. 5, p. 102276, 2020.
16. R. Anand, R. Jestadi, K. Govinda, and D. Sharma, "Data-driven insights: Analyzing and visualizing trends in the chocolate market for strategic decision-making," *AVE Trends in Intelligent Technoprise Letters*, vol. 1, no. 2, pp. 60–70, 2024.
17. H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, and J. Zhu, "Towards multi-pose guided virtual try-on network," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Seoul, South Korea, 2019.
18. M. Vijay, N. Shanmukh, V. S. R. P. Naidu, P. Vinod Kumar, and G. S. S. Narayana, "Enhancing virtual clothing try-on systems: GAN-powered image analysis and generation," in *Proc. 2024 Int. Conf. Recent Innovation in Smart and Sustainable Technology (ICRISST)*, Bengaluru, India, 2024.