Neural Computing and Applications (2025) 37:6125–6144
https://doi.org/10.1007/s00521-024-10843-6

ORIGINAL ARTICLE

Revolutionizing online shopping with FITMI: a realistic virtual try-on solution

Tassneam M. Samy[1] · Beshoy I. Asham[1] · Salwa O. Slim[1] · Amr A. Abohany[2,3]

Received: 23 September 2024 / Accepted: 29 November 2024 / Published online: 9 January 2025

Abstract

In today's digital age, consumers increasingly rely on online shopping for convenience and accessibility. However, a significant drawback of online shopping is the inability to physically try on clothing before purchasing. This limitation often leads to uncertainty regarding fit and style, resulting in customer post-purchase dissatisfaction and higher return rates. Research indicates that online items are three times more likely to be returned than in-store ones, especially during the pandemic. To address this challenge, we propose a virtual try-on method called FITMI, an enhanced Latent Diffusion Textual Inversion model for virtual try-on purposes. The proposed architecture aims to bridge the gap between traditional in-store try-ons and online shopping by offering users a realistic and interactive virtual try-on experience. Although virtual try-on solutions already exist, recent advancements in artificial intelligence have significantly enhanced their capabilities, enabling more sophisticated and realistic virtual try-on experiences than ever before. Building on these advancements, FITMI surpasses ordinary virtual try-ons relying on generative adversarial networks, often producing unrealistic outputs. Instead, FITMI utilizes latent diffusion models to generate high-quality images with detailed textures. As a web application, FITMI facilitates virtual try-ons by seamlessly integrating images of users with garments from catalogs, providing a true-to-life representation of how the items would look. This approach differentiates us from competitors. FITMI is validated using two widely recognized benchmarks: the Dress-Code and Viton-HD datasets. Additionally, FITMI acts as a trusted style advisor, enhancing the shopping experience by recommending complementary items to elevate the chosen garment and suggesting similar options based on user preferences.

Keywords FITMI virtual try-on · Generative architectures · Latent diffusion models

1 Introduction

Online shopping has rapidly expanded in recent years, driven by its convenience and accessibility [28]. However, despite this growing popularity, a significant challenge persists: the inability of customers to physically try on clothing before purchasing [20]. This limitation leads to uncertainties about fit and style, often resulting in higher return rates and decreased customer satisfaction. Existing virtual try-on methods attempt to address these challenges by allowing customers to visualize clothing on a virtual model; however, they predominantly rely on generative adversarial networks (GANs) [13], which can frequently produce low-resolution and unrealistic outputs.

In contrast, recent years have seen the emergence of diffusion models as a leading class of generative architectures for image generation, offering superior quality

✉ Tassneam M. Samy
Tasnim_Mohsen_1375@fci.helwan.edu.eg

Beshoy I. Asham
Bishoy_Ibrahim_1032@fci.helwan.edu.eg

Salwa O. Slim
salwaosama@fci.helwan.edu.eg

Amr A. Abohany
Cio_kfs@kfs.edu.eg

[1] Department of Computer Science, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

[2] Department of Information Systems, Faculty of Computers and Information, Damanhour University, Damanhour, Egypt

[3] Department of Information Systems, Faculty of Computers and Information, Kafrelsheikh University, Kafrelsheikh, Egypt

compared to GANs [3]. Inspired by the success of these generative models, this work introduces FITMI. This virtual try-on solution employs latent diffusion models (LDMs) to overcome the limitations associated with traditional GAN-based methods [26]. LDMs, operating in the latent space of a pre-trained auto-encoder, strike an optimal balance between computational efficiency and image quality, producing high-resolution, realistic images that enhance the online shopping experience.

FITMI took this advancement further by building upon a pre-trained model initially integrated innovative techniques such as the textual inversion network [19] and the Enhanced Mask-Aware Skip Connection (EMASC) module [21] to boost image quality and garment fitting enhancing it with additional capabilities.

These enhancements enable FITMI to generate high-quality, lifelike images with detailed textures, providing users with an accurate and satisfying virtual try-on experience. The FITMI architecture has been validated on two widely-used virtual try-on benchmarks, Dress-Code [23] and Viton-HD [5]. The system's seamless integration mechanism automatically selects the appropriate dataset based on the category argument (upper, lower, or dresses), simplifying the user process and optimizing system performance by applying the most relevant data without manual intervention.

FITMI also incorporates customized preprocessing pipelines for input garments and person images. These pipelines are meticulously designed to meet the specific requirements of the Viton-HD [5] and Dress-Code [23] datasets, ensuring the highest data quality for accurate and realistic model outputs. Moreover, FITMI addresses the challenges posed by diverse body types and complex poses through a comprehensive integration of pose mapping and garment warping techniques. The system employs advanced models, Open Pose [4] and Dense Pose [31], within its preprocessing pipeline to extract detailed pose maps. These maps capture key body landmarks like joints and limbs, ensuring accurate representation of the user's anatomy, regardless of pose complexity. FITMI also includes a garment recommendation system that suggests items based on color and texture similarities of the input garments, alongside a gender-specific recommendation system for complementary items.

By offering a realistic virtual try-on experience, FITMI can significantly benefit apparel businesses, particularly online retailers. It helps customers make more confident purchasing decisions, potentially leading to fewer returns and increased satisfaction. Consequently, FITMI represents a valuable tool for reducing return-related issues and enhancing customers' overall shopping experience.

1.1 Contributions

In summary, this work introduces several key innovations that significantly advance the field of virtual try-on technology:

- *LDM-Based Approach* Unlike traditional methods that predominantly rely on GANs, FITMI employs LDMs. This innovative approach enhances the quality of virtual try-on outputs, producing more realistic and detailed images, thereby improving the user experience and satisfaction.
- *Benchmark Utilization* FITMI has been validated using two widely recognized benchmarks: the Dress-Code dataset [23] for lower-body and dress categories, and the Viton-HD dataset [5] for upper-body categories. These benchmarks provide a robust foundation for evaluating the model's performance across different types of garments and ensuring its generalizability.
- *Automatic Dataset Integration* The system features a seamless dataset integration mechanism that automatically selects the appropriate dataset based on the category argument (upper, lower, or dresses). This functionality eliminates the need for manual selection, streamlines the workflow, and enhances the user experience by ensuring the most relevant data is applied.
- *Custom Preprocessing Pipelines* FITMI incorporates customized preprocessing pipelines designed explicitly for input garments and person images. These pipelines are meticulously tailored to meet the distinct requirements of the Viton-HD [5] and Dress-Code [23] datasets, ensuring high-quality data input, which is critical for generating accurate and realistic virtual try-on results.
- *Advanced Functionalities* FITMI introduces advanced functionalities, including a garment recommendation system that suggests items based on color and texture similarities. Additionally, it incorporates a gender-specific recommendation system for complementary items, significantly enhancing the user experience by providing personalized and context-aware shopping suggestions.

1.2 Paper structure

The remainder of this paper is structured as follows: Sect. 2 reviews the existing research relevant to our study, providing essential background information. In Sect. 3, we describe FITMI's proposed solution, detailing the system's architecture, including the preprocessing techniques and the models employed. Section 4 discusses the datasets

used, the experimental setup, technical aspects, and evaluation metrics. This section also presents the testing and evaluation processes, the results of the proposed model, and the development and deployment of the FITMI web application. Section 5 provides a comprehensive discussion on the methodology behind FITMI's development, focusing on the strategic selection of datasets and the different preprocessing approaches utilized. Finally, Sect. 6 summarizes the key findings of FITMI, draws conclusions, and outlines potential avenues for future research.

2 Related work

Significant advancements and methodologies have been developed to improve virtual try-on systems, enhancing accuracy, realism, and user experience. This section provides an overview of image-based virtual try-on methods that utilize cloth deformation, emphasizing the preservation of clothing details through explicit warping modules that adapt the garment to a person's body. The discussion then shifts to segmentation generation for try-on synthesis, highlighting the need to isolate body and garment parts accurately for seamless integration. The generation phase and refinement processes are also examined to understand how high-quality try-on images are produced, although some methods still result in low-resolution outcomes. Finally, parser-free methods are reviewed as an alternative to streamline the virtual try-on process by bypassing detailed parsing of user images while maintaining or improving visual quality and accuracy. The section concludes by exploring targeted solutions addressing these limitations, focusing on enhancing visual coherence and garment realism.

2.1 Clothes deformation

Many virtual try-on systems aim to transfer a desired garment onto a target subject's corresponding region while preserving the clothing item's intricate details.

To maintain these details, earlier approaches utilized explicit warping modules, which adapt the input garment to the body shape of the target person. A notable example is the pioneering work Virtual Try-On (VITON) [11], which introduced a framework consisting of an encoder–decoder generator. This generator produces a coarse virtual try-on result, which is then refined by a network that applies thin-plate spline transformation [8] to warp the clothing to fit the target subject.

However, despite improvements in warping techniques, misalignment between the warped clothes and the subject's body remains a significant issue, often leading to visual artifacts in misaligned regions. Additionally, VITON and similar methods lack the integration of person representation, fail to handle pose variations, and do not account for clothing deformations caused by movement. As a result, these approaches struggle to accurately position body parts like the arms and hands, ultimately preventing the generation of fully photo-realistic virtual try-on results.

2.2 Segmentation generation for try-on synthesis

High-resolution virtual try-on methods [5, 16, 32] generally include a segmentation generation module because the importance of precise segmentation maps increases with image resolution. The segmentation map is essential for correctly delineating body parts and garments, ensuring seamless integration of clothing onto the model. Recently, Viton-HD [5] introduced a normalization technique to reduce misalignment issues, though this method struggles to fill misaligned regions with appropriate clothing textures.

A commonality among these techniques is their reliance on GANs [10, 16] for image generation during the try-on process. In High-Resolution Virtual Try-On (HR-VITON), Lee et al. [16] tackled the misalignment issue by developing a unified pipeline that integrates warping and segmentation stages, resulting in more precise high-resolution outcomes.

Another approach, Adaptive Content Generating and Preserving Network (ACGPN) [32], focuses on masking the clothing region in the person's image and reconstructing it with the target clothing. This method relies heavily on accurate human parsing for garment placement. Still, it tends to lose the unique characteristics of the clothing during training, as it overemphasizes the silhouette of the original garment.

Unfortunately, methods that depend on human parsing are prone to errors-minor mistakes in segmentation can result in unrealistic try-on images with significant artifacts. These issues underscore the need for further refinement in segmentation techniques to preserve garment texture and fit while reducing artifacts and improving realism.

2.3 Generation phase and refinement of the result

Another critical research direction involves the generation phase and the subsequent refinement of results in virtual try-on systems. In their work on the Dress-Code dataset, Morelli et al. [23] concentrated on enhancing the semantic understanding of generated images. A semantic-aware discriminator was introduced, which operates at the pixel level, enhancing the realism and visual quality of the output. By learning the internal representation of semantic

content within images, this method effectively captures the nuances of clothing textures, fabric movement, and body integration, improving the fidelity of the virtual try-on experience.

Despite these advancements, limitations persist—particularly when the system generates low-resolution or less detailed results, which can detract from the overall realism. Low-resolution images struggle to capture fine details, such as intricate textures or subtle fabric behaviors, essential for a convincing virtual try-on.

Further refinement and optimization in virtual try-on systems are urgently needed to address these issues. Key challenges include balancing computational efficiency with generating high-quality, detailed images. As resolution and complexity increase, the demand for more powerful generative models grows, necessitating image synthesis and refinement techniques advancements.

Overcoming these obstacles is crucial for making virtual try-on systems more broadly applicable, particularly in commercial settings. Continued research and development are essential to enhance virtual garment simulations' realism, accuracy, and overall effectiveness, paving the way for their widespread adoption in e-commerce and fashion technology.

2.4 Parser-free methods

Pioneering parser-free methods [10, 18] have made significant advancements in generating high-quality outputs without relying on traditional parsing techniques typically used to segment the human body and clothing. These approaches eliminate the need for detailed human parsing, simplifying the virtual try-on process and speeding up computations. However, despite their potential, early parser-free methods still encounter several challenges.

One notable issue arises from using the same input–output pairs for both Teacher and Student networks, which can result in artifacts in the generated outputs. To mitigate this problem, Ge [10] introduced the Parser-Free Virtual Try-on via Distilling Appearance Flows (PF-AFN), a novel Knowledge Distillation-based training pipeline. In this method, the Student network is trained using the Teacher network's output as input, while the original images are used to supervise the Student network's output directly. This approach has set a new benchmark for parser-free methods, significantly enhancing the generated results' quality, coherence, and accuracy.

Building on these advancements, A Regional Mask Guided Network for Parser-free Virtual Try-on (RMGN) [18] further refined the generation process by integrating SPADE blocks [24], which allow for spatially adaptive denormalization. This enhancement improved the system's ability to retain details and produce more realistic images, mainly when dealing with complex textures and clothing patterns.

Despite these improvements, many parser-free methods still depend on parser-based approaches, particularly for tasks like pose estimation or garment alignment. Human representation parsing continues to be a time-consuming and computationally expensive process. Optimizing these aspects remains a crucial challenge for further improving the efficiency and accuracy of parser-free virtual try-on systems, ensuring they can achieve the same level of detail and realism as parser-based methods.

2.5 Damage detection and optimization

Advanced optimization techniques, such as particle swarm optimization (PSO) and hybrid algorithms [14], have the potential to significantly enhance feature extraction in garment fitting applications. PSO, inspired by the social behavior of birds, iteratively optimizes candidate solutions to improve accuracy in predictions. This technique can be utilized to fine-tune hyperparameters of machine learning models or to select the most relevant visual features-such as edges, colors, and textures—that influence how garments appear on different body types. Hybrid algorithms, which combine the strengths of multiple optimization methods, can further improve the model's ability to recognize complex patterns in garment images and user poses, leading to more precise fitting predictions.

Incorporating dynamic analysis and finite element modeling (FEM) [15] can also refine and simulate how garments behave when worn. Dynamic analysis can capture the movement of garments during activities like walking or bending, allowing for a more realistic representation in virtual try-ons. FEM can model interactions between garments and various body types, providing insights into how fabric drapes and flows, ultimately enhancing the prediction of fit and appearance.

2.6 Overcoming limitations in virtual try-on systems

Previous virtual try-on systems have faced numerous challenges that hinder their performance and realism. Traditional methods, such as those that rely on explicit warping modules like VITON [11], often suffer from misalignment issues, leading to noticeable artifacts in the final output. These methods typically struggle to accurately position body parts such as the arms and hands, resulting in unrealistic, less immersive experiences for users. Additionally, segmentation-based approaches, such as VITON-HD [5] and ACGPN [32], encounter difficulties, particularly in filling misaligned regions with proper clothing textures. Their heavy reliance on accurate human parsing

means that even minor errors in segmentation can result in significant artifacts, which degrade the overall quality of the generated images.

While parser-free methods aim to simplify the try-on process by eliminating the need for detailed human parsing, they introduce their own set of challenges. These methods often struggle with artifacts arising from repetitively using the same input–output pairs during training. Despite reducing the complexity of parsing, these approaches still rely heavily on parsing processes for other stages, making them time-consuming and error-prone.

To address these limitations, our proposed system, FITMI, integrates LDMs and advanced image–text fusion techniques, which offer significant improvements over traditional methods. By leveraging LDMs with dual inputs-the target garment and the model's pose-FITMI ensures precise alignment of garments with the model's body, eliminating the misalignment and artifact issues seen in earlier approaches.

One of FITMI's key innovations is the EMASC module, which preserves high-frequency details during image generation. This allows FITMI to accurately render even complex garment textures, such as intricate patterns or delicate fabrics. Furthermore, by incorporating Stable Diffusion techniques, FITMI can seamlessly fill misaligned regions, enhancing the visual coherence and realism of the generated images.

Another crucial aspect of FITMI's architecture is integrating a text time-conditional U-Net and CLIP text encoder, which significantly improves the precision of garment details. These components help address low-resolution issues that have traditionally plagued virtual try-on systems, resulting in high-quality, photo-realistic images that accurately represent the garments.

FITMI's adoption of the LaDI-VTON model [22], the first latent diffusion-based method for virtual try-on, further enhances its effectiveness. By utilizing this cutting-edge approach, FITMI has outperformed its competitors in widely recognized benchmarks, such as the Dress-Code [23] and Viton-HD [5] datasets. These advancements have enabled FITMI to achieve superior realism and effectiveness in generating high-quality virtual try-on experiences.

By also integrating hybrid machine learning models, pose estimation, and dynamic analysis techniques even though dynamic analysis traditionally focuses on systems in motion, it can be adapted for images by using them in conjunction with motion data or simulations. This can create a more realistic understanding of how garments behave during wear, ultimately leading to better fitting, and FITMI can further optimize its feature extraction process, enabling it to capture finer garment details and deliver a more accurate and satisfying virtual try-on experience

3 Proposed method

Our proposed system, as shown in Fig. 1, introduces a novel approach to virtual try-on by integrating LDMs with advanced image–text fusion techniques. The system leverages Stable Diffusion, an LDM architecture known for its superior image generation capabilities compared to traditional GANs. Building on this foundation, we have introduced several enhancements designed for virtual try-on scenarios.

A key feature of our approach is the use of dual inputs-the target garment and the model's pose—which enables the system to generate highly detailed and realistic depictions of how the garment fits. This dual-input design ensures that the virtual try-on maintains the model's physical characteristics, pose, and identity, accurately representing the garment to the user.

To enhance precision, a text time-conditional U-Net denoising model is incorporated besides a CLIP text encoder, significantly refining the garment's details and textures. This integration ensures that the delicate nuances of fabric, color, and pattern are captured accurately, which is crucial for maintaining fidelity in virtual try-on applications.

Further, our system includes an EMASC module, which addresses the challenge of retaining high-frequency details—such as fine textures and small garment features—during the image reconstruction phase. By preserving these intricate features, the EMASC module enhances the realism and overall quality of the virtual try-on experience.

In addition to these technical components, the FITMI application is designed with user-centric features that enhance the virtual try-on experience. By leveraging advanced machine learning algorithms, FITMI offers personalized garment recommendations based on user preferences. Using the Dress-Code dataset [23], which includes categories such as upper-body, lower-body, and dresses, garment embeddings are generated using a pre-trained ResNet50 convolutional neural network (CNN). This CNN extracts essential image features, and the resulting embeddings are normalized and stored, allowing the system to provide accurate and personalized recommendations.

Moreover, FITMI extends beyond single-item garment suggestions by offering complementary item recommendations. These recommendations help users complete their look with curated accessories, shoes, or outerwear, catering to a diverse audience across all genders. For this feature, complementary item embeddings were generated using the same process as garment recommendations. However, datasets such as the Real Fashion dataset [1]—focusing on accessories, footwear, and outerwear—were incorporated.

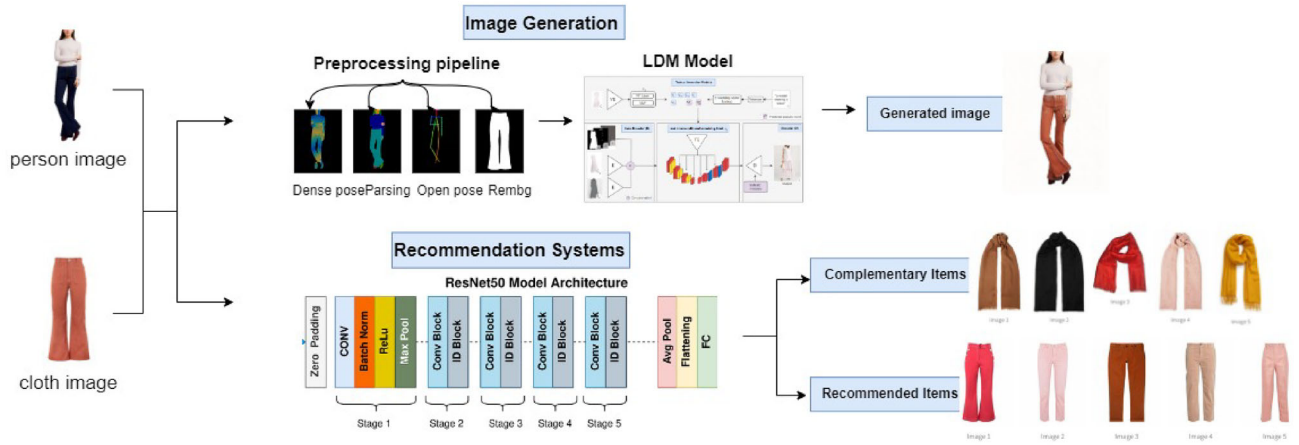

Fig. 1 Overview of the proposed system

Additionally, gender-specific datasets were assembled, ensuring the system captures style preferences and trends for both male and female users, allowing for a more personalized and comprehensive virtual shopping experience.

3.1 Preprocessing

Preprocessing is a foundational stage within our methodology, as our proposed model relies on it to accurately classify and recognize patterns in the datasets. The preprocessing workflow, as shown in Fig. 2, orchestrates a sequence of four distinct models—Mask and Background Removal, Parsing, Open Pose, and Dense Pose—each integrated to execute consecutively. This step-by-step process ensures the input data is clean, segmented, and accurately mapped for optimal results in the virtual try-on system.

3.1.1 Background removal mask

To enhance the quality of our results, a background removal mask strategy has been implemented by utilizing images with a white background. This deliberate choice has consistently produced superior outcomes in the virtual try-on process. To streamline this aspect of our pipeline, the Remove Background (Rembg) module [9] has been integrated. This powerful library specializes in background removal from images, providing a robust solution for our preprocessing needs.

As shown in Fig. 3, the Rembg module is critical in ensuring clean data by removing unwanted backgrounds. The optional 'mask' parameter within the Rembg module is leveraged to isolate specific areas or objects in an image, ensuring the precise extraction of both the garment and the person from their backgrounds. This strategic approach improves the clarity and accuracy of the input data, preparing it for subsequent stages in the workflow.

3.1.2 Parsing

Parsing, also known as segmentation or label mapping, is a technique used to partition the human body into distinct semantic regions, covering essential areas such as the head, torso, arms, legs, and finer details like hair and clothing. In our methodology, we employed the Self Correction for Human Parsing model [17] was trained on three diverse datasets to ensure comprehensive segmentation.

Fig. 2 Preprocessing pipeline

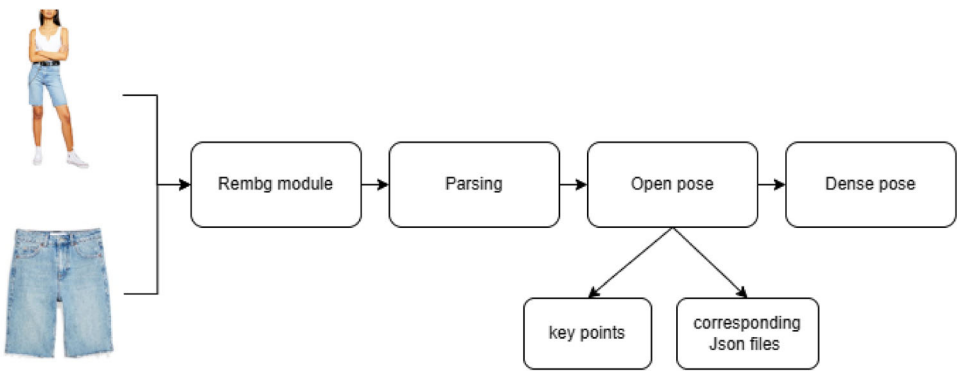

Fig. 3 Rembg module

Fig. 4 Parsing model

As shown in Fig. 4, the model divides the body into relevant fashion-related semantic categories. Two datasets, in particular, were well-suited for our purposes: the ATR and LIP datasets. The ATR dataset offers 18 semantic category labels, including vital elements like 'face,' 'sunglasses,' 'hat,' 'scarf,' 'hair,' 'upper clothes,' 'pants,' and 'shoes.' Meanwhile, the LIP dataset extends this further with 20 categories, adding labels such as 'Coat' and 'Jumpsuits.'

Focusing on these fashion-centric categories, our parsing methodology deeply understands garment-related features, ensuring accurate placement and segmentation for virtual try-on purposes.

3.1.3 Open Pose

Open pose provides accurate human pose estimation by offering detailed insights into critical body joints. In our framework, the primary goal of Open Pose is to ensure the model's body pose is faithfully preserved throughout the processing stages, which is vital for accurate virtual try-on results.

To achieve this, we integrated the PyTorch implementation of the Open Pose model [4] due to its compatibility with our workflow and ability to extract essential pose information. As shown in Fig. 5, the Open Pose model efficiently captures two key outputs:

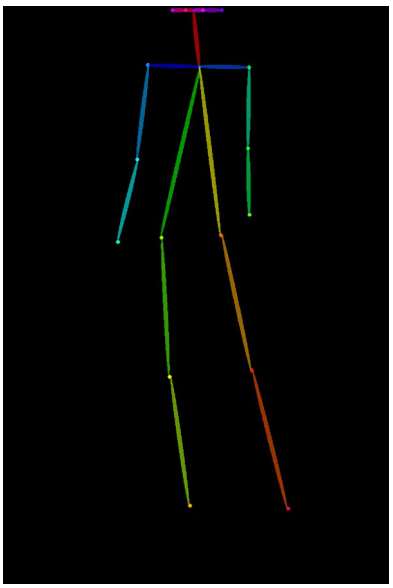

Fig. 5 Open pose model

1. *Key points* that indicate pose estimation.
2. *JSON files* that store the precise coordinates of these key points.

These key points form the basis for computing an *18-channel pose heatmap*, where each channel corresponds to a distinct body joint, such as the shoulders, elbows, or knees. This heatmap comprehensively represents the subject's pose, enabling our preprocessing pipeline to ensure accurate body pose representation for garment fitting.

3.1.4 Dense Pose

Dense Pose strives to establish dense correspondence between every pixel within an image and specific points on the human body surface. This intricate mapping facilitates a pixel-to-surface correlation, wherein each pixel is assigned a unique body surface identifier and local coordinates, yielding a comprehensive understanding of the body's shape and appearance as shown in Fig. 6. Given the

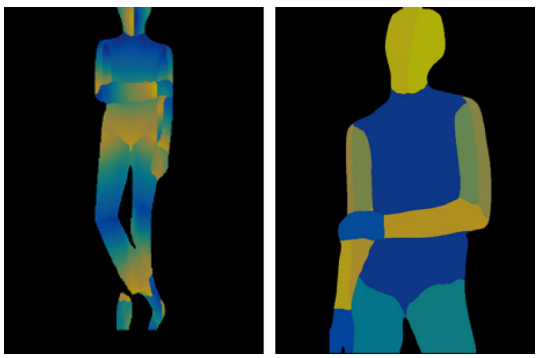

Fig. 6 Dense Pose model

computational demands and processing time associated with Dense Pose, integrating with the detectron2 library [31] was opted. Leveraging detectron2's capabilities, estimating the U-coordinates for various body parts was a focus. This strategic choice enables us to accurately represent body part positioning while mitigating the computational overhead associated with Dense Pose.

The output of Dense Pose comprises two critical components: the 25-channel label map and the 2-channel UV map. These maps encapsulate invaluable information regarding body part segmentation and surface coordinates. In our preprocessing pipeline, these maps were concatenated without further processing.

In conclusion, our preprocessing pipeline orchestrates a series of sophisticated techniques; each is used to enhance the quality and relevance of the data input into our proposed model. The garment and person were extracted from the background, beginning with the Mask Model. Parsing extracts semantic information about the human body and attire. Open Pose and Dense Pose offer detailed insights into body pose and surface correspondence to the body's shape and appearance. By integrating these components seamlessly, our preprocessing pipeline ensures comprehensive data extraction, empowering our model to make accurate classifications. This precisely crafted workflow optimizes accuracy and fosters robustness and efficiency in our methodology.

3.2 Model

FITMI's approach innovatively integrates LDMs with the latest image and text integration advancements, drawing inspiration from successful precedents and extending their capabilities. The backbone of our method lies in adopting LDMs, which have been shown to surpass the capabilities of traditional GANs in image generation tasks. This shift was initially inspired by the work presented by Rombach et al. [27], which mitigated some of the inherent limitations of GANs. The LaDI-VTON model [22] adapted LDMs by effectively using latent space for more detailed and controllable image synthesis.

The standard input structure of these models has been modified to accommodate the complexities of virtual try-on better. Moving beyond the typical single-input systems that focus solely on the garment, our FITMI model uniquely incorporates dual inputs: the garment itself and the pose of the target model. This enables a more detailed depiction that not only conforms the garment to the model but also precisely aligns it with the model's pose, ensuring that the physical characteristics, pose, and identity are convincingly maintained.

3.2.1 Latent diffusion models 'Stable Diffusion'

Understanding the underlying components of Stable Diffusion, an LDM architecture, is essential. The model comprises an auto-encoder with distinct encoder and decoder components, a text time-conditional U-Net denoising model, and a CLIP text encoder. Figure 7 illustrates the model used in our proposed system in Fig. 1.

The Stable Diffusion model integrates visual and textual data to generate high-quality images. The key components include:

- Visual Encoder (VE) converts the clothing image into a visual embedding.
- Text Encoder (TE) processes textual descriptions to produce text embeddings.
- Denoising U-Net that iteratively refines a concatenated representation of visual and text embeddings, ensuring the generated image aligns with the textual description.

The process begins with encoding the clothing image and text. The auto-encoder encodes images of the clothing item and a model wearing plain clothes into latent representations. These are then concatenated with visual and textual embeddings. The U-Net denoises this combined representation, which the decoder transforms into a realistic image of a model wearing the dress. This enhances the virtual try-on experience by providing detailed and accurate visualizations.

3.2.2 Text time-conditional U-Net denoising model paired with a CLIP text encoder

Central to the FITMI method is the refined use of a text time-conditional U-Net denoising model paired with a CLIP text encoder within the diffusion model framework. The denoising network, pivotal in the diffusion process, minimizes a loss function associated with reconstructing images deliberately corrupted with Gaussian noise. Furthermore, the CLIP model, a vision-language model integral to the textual inversion technique, aligns visual and textual data within a unified embedding space where visual features of the garment are encoded as textual token embeddings. These are pseudo-word token embeddings since they do not correspond to any linguistically meaningful entity but represent the visual features of the in-shop garment in the token embedding space.

Such an arrangement allows the model to accelerate diffusion by focusing on these tokens, facilitating a more seamless integration of text and image features. This approach enhances the precision of the garment's details and textures in the final output, ensuring the integrity and fidelity of complex textile patterns and colors in virtually tried-on clothing.

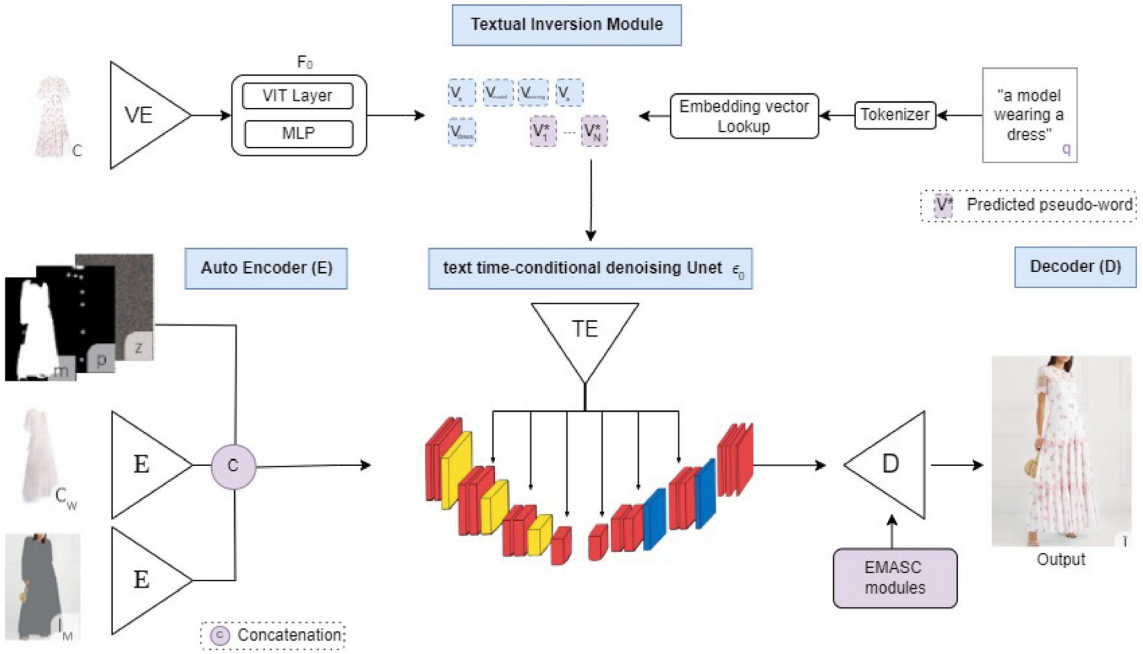

Fig. 7 Overview of the proposed model

3.2.3 Enhanced mask-aware skip connection module

As we discussed, LDMs are augmented with a textual inversion network. However, LDMs encounter difficulties with fine details like high frequency in the pixel space due to the spatial compression performed by the auto-encoder. Consequently, to address the challenges of high-frequency detail retention in LDMs, particularly in areas like hands, feet, faces, and regions where the new garment is incorporated, the EMASC module has been implemented. The EMASC module specializes in inpainting by substituting garment information in human-based images with details from a target garment image provided by the user.

EMASC is an architectural improvement that significantly enhances the model's performance. Notably, masked skip connections are integrated within the auto-encoder structure of the Stable Diffusion model. These connections are crucial for maintaining high-quality image outputs, as they play a key role in preserving details during the image reconstruction phase. This enhancement improves the model's ability to transfer intricate information from the encoding phase back to the decoding phase.

The effectiveness of FITMI's approach has been validated against prominent virtual try-on benchmarks such as Viton-HD [5] and Dress-Code [23]. Our method demonstrates superior quantitative and qualitative performance compared to existing state-of-the-art techniques.

4 Experimental results

The experimental results are evaluated based on several components. The datasets utilized include Dress-Code [23] and Viton-HD [5], which provide the foundation for assessing the model alongside the experimental setup that ensures reproducibility. The evaluation metrics employed are Kernel Inception Distance (KID), Fréchet Inception Distance (FID), Structural Similarity Index (SSIM), and the Learned Perceptual Image Patch Similarity (LPIPS), which are used to assess various aspects of the model's performance. This is followed by a presentation and analysis of the results obtained from the FITMI proposed model.

4.1 Datasets

Due to the strategic role that virtual try-ons play in e-commerce, many rich and potentially valuable datasets are proprietary and not publicly available to the research community. Public datasets, instead, either do not contain paired images of models and garments or feature a minimal number of images. Moreover, the overall image resolution is low, mostly (256 $\times$192). Unfortunately, these drawbacks slow down progress in the field. Our architecture is validated on two widely used virtual try-on benchmarks (i.e., Dress-Code [23] and Viton-HD [5]), as it outperforms competitors in terms of realism across both datasets. Images from both datasets exhibit great variety, considering the reference models' body poses and the categories and textures of try-on garments. This helps virtual try-on

architectures become more general and adapt to challenging scenarios.

Viton-HD [5] is a (1024 × 768) virtual try-on dataset that contains only upper-body clothes with 13,679 frontal-view woman and top clothing image pairs. The pairs are split into a training and test set with 11,647 and 2032 pairs, respectively. One can either use the pairs of a person and a clothing image to evaluate a paired setting or shuffle the clothing images for an unpaired setting. The paired setting aims to reconstruct the person's image with the original clothing item, while the unpaired setting changes the item on the person's image with a different item.

It utilizes seven techniques, as shown in Fig. 8, including masks to differentiate the person from the background, Dense Pose, and Open Pose for detailed pose estimation, and parsing representations for segmenting the body into distinct parts. The system also incorporates agnostic masks and parsing to focus on body shapes rather than specific clothing, providing an agnostic representation of the person that abstracts away physical and clothing details using pose and segmentation maps. Additionally, a particular mask is used for the clothing items.

Dress-Code [23] is more than 3× larger than publicly available datasets for image-based virtual try-ons and features high-resolution paired images (1024 × 768) with front-view, full-body reference models. It contains more than 50,000 image pairs of try-on garments and corresponding catalog images, where a model wears each item. A dataset of this scale is preferable over other datasets with similar characteristics but a petite size, as it provides more data for training and validation. Dress-Code is the first publicly available dataset featuring lower-body and full-body clothes. All images are in high resolution (1024 × 768), making it more than 3× larger than Viton-HD, which contains only upper-body garments.

Dress-Code features three categories:

- Upper-body (composed of tops, T-shirts, shirts, sweatshirts, and sweaters),
- Lower-body (composed of skirts, trousers, shorts, and leggings),
- Full-body clothes (composed of dresses).

The dataset comprises 53,795 image pairs: 15,366 for upper-body clothes, 8951 for lower-body clothes, and 29,478 for dresses. To preserve the models' identity, all images are partially anonymized by cropping at the nose level. This ensures that no information about the human models' countenance is available.

Dress-Code [23] contains images paired with corresponding clothing items that undergo six preprocessing steps, four of which are shown in Fig. 9. The six preprocessing steps include joint coordinates in (.npy) loading arrays, which denote critical points on the body for pose estimation. The dataset also contains dense pose information, which maps image pixels to the 3D surface of the human body for detailed pose understanding. Additionally, it offers cloth align masks and cloth align parses, which assist in aligning the clothing properly on the body. Cloth-warped images depict the clothing items adjusted to fit the body shape. Finally, the dataset provides parsing information about the person, segmenting the image into different semantic parts for all three categories of clothing: upper, lower, and dresses.

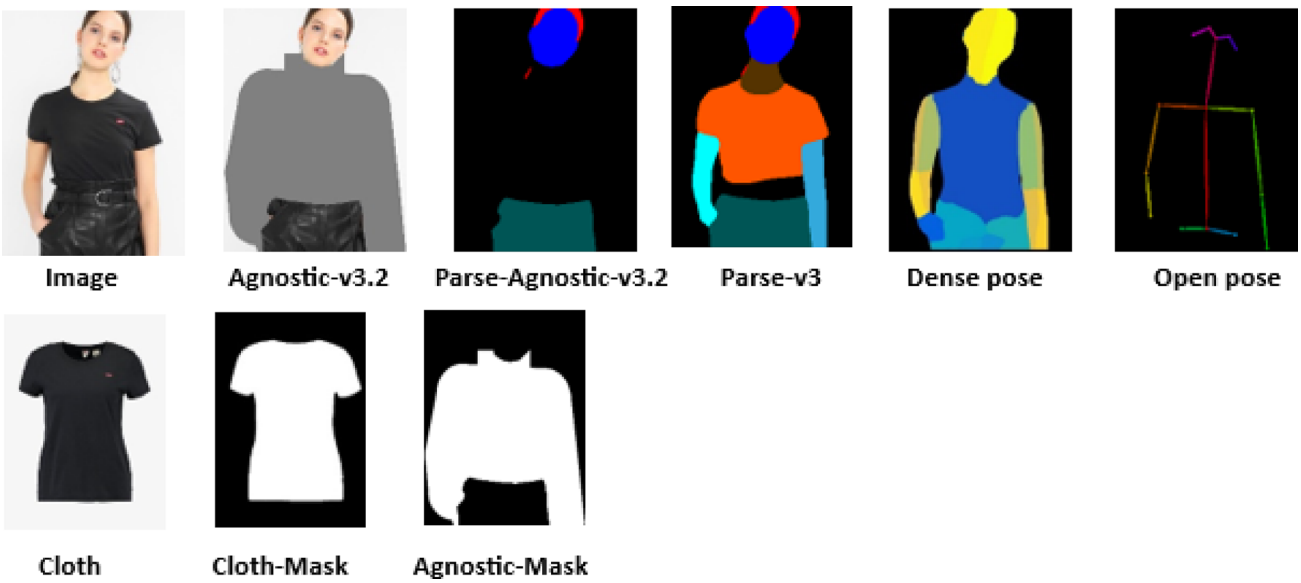

Fig. 8 Viton-HD dataset categories

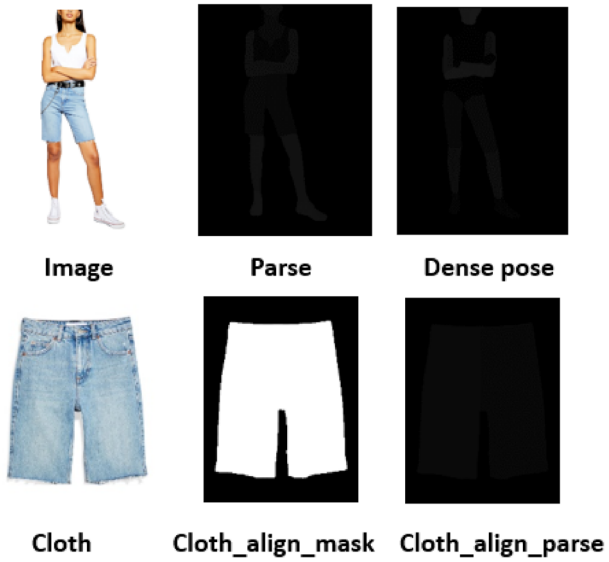

Fig. 9 Dress-Code dataset categories

4.2 Experimental setup

Our implementation utilizes the LaDI-VTON pre-trained model [22], which comprises three primary modules: EMASC, textual inversion adapter, and warping component. The pre-trained model is employed for testing purposes without any additional training. Weight freezing is applied to all modules except for the textual inversion adapter, which is evaluated alongside the enhanced Stable Diffusion pipeline. Image generation occurs at a resolution of 512×384 pixels. The textual inversion network F_θ features a single Vision Transformer (ViT) layer followed by a multilayer perceptron (MLP) with three fully connected layers. It generates 16 PTEs, with training configurations including 200k training steps, a batch size of 16, a learning rate of 1×10^{-5} (with 500 warm-up steps), and the AdamW optimizer set with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1×10^{-2}. The diffusion virtual try-on model is tested under predefined optimizer and scheduling strategies. Input components are randomly masked with a probability of 0.2, and the model is trained for 200k iterations using settings similar to those of the textual inversion network. For the auto-encoder with EMASC, the configuration includes two convolutional layers with a SiLU nonlinearity. These modules are trained for 40k steps, also utilizing a batch size of 16, a learning rate of 1×10^{-5}, and the AdamW optimizer with the same hyperparameters as mentioned previously. Loss functions applied during testing include a combination of L1 and VGG loss functions, while the encoder and decoder remain frozen throughout testing, focusing the assessment solely on the EMASC modules.

Enhancing the proposed model, a preprocessing pipeline was customized to accommodate the specific requirements of the two datasets used: Dress-Code [23] for lower-body and dresses and Viton-HD [5] for upper-body. Both datasets undergo a standardized preprocessing pipeline, differing only in parsing and dense pose methodologies. For the Dress-Code dataset [23], parsing is facilitated by the Self Correction for Human Parsing model [17], pre-trained on the ATR dataset that boasts 18 semantic category labels, while the Viton-HD dataset [5] parsing model is pre-trained on the LIP dataset, which provides 20 semantic category labels.

Notably, a divergence occurs in utilizing Dense Pose detectron2 [31] parameters. In the Viton-HD dataset [5], the `dp_segm` parameter is employed to generate segmentation masks for annotated persons, optimizing pose estimation accuracy. Conversely, in the Dress-Code dataset [23], the `dp_u` parameter is utilized for point annotation colored according to their U coordinate in part parameterization.

Despite parsing and dense pose variations, all other models in the pipeline remain consistent for both datasets. The input garment image and the person image both traverse through the Rembg module [9], followed by parsing [17], Open Pose [4], and Dense Pose [31] models sequentially. Subsequently, they are fed into the pre-trained model, generating the final output of the person wearing the new desired garment.

For our features, a recommendation system was implemented for fashion items based on garment input texture and color. Additionally, a complementary items recommendation system was implemented based on gender. Both systems employ advanced deep learning techniques for feature extraction and similarity search to provide personalized recommendations to users. It begins by loading a pre-trained ResNet50 model. The model is initialized with weights pre-trained on the ImageNet dataset [6] and is configured to exclude the top classification layer, ensuring that it captures high-level features relevant to garment images.

Additionally, for computing embeddings and filenames, a custom dataset is employed that integrates data from three distinct sources: Viton-HD [5], Dress-Code [23], and Real Fashion [1] datasets for each system, consisting of garment and accessories images. Creating embeddings from these datasets involved several steps, allowing for efficient feature retrieval during recommendation. Garment images are preprocessed and passed through the pre-trained ResNet50 model to extract high-dimensional feature vectors. These feature vectors encode semantic information about the garments, capturing their visual characteristics in a compact representation. The extracted features are normalized to unit length to ensure consistency and

comparability. The recommendation system employs a nearest neighbors algorithm to find similar garment images based on the extracted features. By calculating the Euclidean distance between feature vectors, the system identifies garments most visually identical to the input image. The number of nearest neighbors and the distance metric are configurable parameters that can be adjusted to fine-tune the recommendation results to deliver personalized and visually appealing fashion recommendations to users.

4.2.1 Computational resource requirements

Since the focus was solely on enhancing an existing pre-trained model, no additional training was necessary. The system was tested using a T4 GPU on Google Colab. However, locally, NVIDIA's CUDA for GPU acceleration was utilized, allowing us to harness the computational power of NVIDIA GPUs effectively. In conjunction with cuDNN, a GPU-accelerated library for deep neural networks, significant performance improvements for inference were achieved. Additionally, CuPy, which provides a NumPy-like interface for GPU computing, facilitated efficient array operations on the GPU, optimizing processing speeds during local testing. Finding compatible versions of these tools was essential to ensure seamless integration within the system. By implementing these enhancements, FITMI optimizes processing efficiency while maintaining high-quality image output. However, high GPU power is recommended for optimal performance, given the demands of the virtual try-on process.

4.3 Evaluation metrics

Experiments on the two virtual try-on datasets Dress-Code [23] and Viton-HD [5] were performed, which feature high-resolution image pairs of in-shop garments and model images in unpaired settings. In the unpaired setting, a garment is selected from the model for the virtual try-on task.

Evaluation metrics were employed to estimate the coherence and realism of the generation to evaluate our model quantitatively. The LPIPS [33] was used to extract feature representations from a pre-trained neural network (e.g., AlexNet or VGG) for both generated and ground truth images. The perceptual distance is computed by comparing these features between corresponding image patches, with lower LPIPS scores indicating higher perceptual similarity. The SSIM [30] was used to evaluate the coherence of the generated image compared to the target image by assessing image quality and decomposing images into local patches, and evaluating luminance, contrast, and structural similarities, yielding a score between –1 and 1, where 1 signifies perfect structural alignment. These metrics were computed using the unpaired settings of both datasets.

To measure realism, the FID [12] was employed to calculate the distance between the mean and covariance of features extracted from generated and lifelike images using the formula:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

where μ_r and Σ_r are the mean and covariance of real image features, and μ_g and Σ_g are the mean and covariance of generated image features. Lower FID values indicate that the generated images are closer to the real images. In unpaired settings, the KID [2] uses features extracted from an Inception network and measures the distance between their distributions with kernel methods. The KID score is averaged over multiple kernels and sample sizes using the formula:

$$\mathrm{KID} = \frac{1}{M}\sum_{i=1}^{M}\left[\mathbb{E}_{x,x'}[k(x,x')] - \mathbb{E}_{x,x''}[k(x,x'')]\right]$$

where k is the kernel function, and x, x', and x'' are samples from the distributions. Lower KID values indicate better alignment with real image distributions. Together, these metrics provide a thorough assessment of image coherence and realism. For the LPIPS and SSIM implementation, the torch-metrics Python package [7] was used, while for the FID and KID scores, the implementation in [25] was employed.

The Viton-HD dataset [5] comprises 13,679 image pairs, each composed of a frontal view of a woman and an upper-body clothing item with a resolution equal to 1024×768. The dataset is divided into training and test sets of 11,647 and 2032 pairs, respectively. Table 1 shows the quantitative analysis of the Viton-HD dataset [5]. FITMI outperforms other competitors significantly in terms of FID and KID, showcasing its effectiveness in this context.

The Dress-Code dataset [23] features over 53,000 image pairs of clothes and human models wearing them. The dataset includes high-resolution images (i.e., 1024 × 768) and garments belonging to different macro-categories, such as upper-body clothes, lower-body clothes, and dresses. In

Table 1 Quantitative results on the Viton-HD dataset [5]

Model	LPIPS ↓	SSIM ↑	FID u ↓	KID u ↓
CP-VTON [29]		0.791	30.25	40.12
ACGPN [32]		0.858	14.43	5.87
Viton-HD [5]	0.116	0.863	12.96	4.09
HR-VTON [16]	0.097	0.878	13.06	4.72
FITMI	0.091	0.876	9.41	1.60

our experiments, the original splits of the dataset were employed, where 5400 image pairs (1800 for each category) compose the test set, and the rest compose the training set.

Table 2 presents the quantitative results on the Dress-Code dataset [23]. As shown, FITMI achieves comparable results in terms of coherence with the inputs (i.e., LPIPS and SSIM) and surpasses all competitors in terms of realism in unpaired settings.

As a complement to Tables 2, 3 presents the complete quantitative results for each category of the Dress-Code dataset [23]. Our method, FITMI, demonstrates superior performance compared to all competitors across all three Dress-Code categories regarding realism metrics such as FID and KID in unpaired settings. Our approach achieves better results when assessing input adherence metrics such as LPIPS and SSIM than Clothing Shape and Texture Preserving Image-Based Virtual Try-On (CP-VTON) [29].

4.4 Results of the proposed model

The qualitative results of the FITMI virtual try-on system showcase its ability to simulate various types of garments on users realistically, with each try-on session taking approximately 10 s per image. In Fig. 10, the top row of images demonstrates the system's capability to handle lower-body garments, including different styles of pants. The bottom row focuses on the upper-body and dress garments, capturing the intricate details and flow of the fabric. These results illustrate our virtual try-on system's high fidelity and versatility, providing users with a true-to-life representation of how clothing items will look when worn.

The input garment and person undergo our preprocessing pipeline, as illustrated in Fig. 2, which are crucial inputs for the model to generate the final results. Additionally, the model features two recommendation systems: one suggesting alternative items based on the input garment and another providing gender-specific complementary items. These features enhance the user experience by offering relevant garment suggestions.

Table 2 Quantitative results on the Dress-Code dataset [23]

	All			
Model	LPIPS ↓	SSIM ↑	FID u ↓	KID u ↓
PF-AFN [10]	–	–	–	–
HR-VTON [16]	–	–	–	–
CP-VTON [29]	0.186	0.842	31.19	25.17
FITMI	0.064	0.906	6.48	2.20

Figures 11, 12, and 13 present the outputs of our recommendation system, highlighting the suggested garments for upper-body, lower-body, and dress categories. The recommendations are generated based on the selected input garment, with complementary items tailored to the user's gender.

4.5 Managerial implications

An aspect of FITMI was the development of a web application to facilitate user interaction with our virtual try-on system. This application, built using Streamlit, provides an intuitive interface for users to interact with and visualize the results generated by our models. The computational demands of this project exceeded the capabilities of standard laptops and desktops, necessitating the use of Google Colab for all computational tasks. Google Colab offers a powerful cloud-based platform with high-performance GPUs, which is crucial for managing the substantial computational load associated with our experiments. This choice ensures that our results are replicable by users who do not have access to specialized hardware, thus democratizing access to high-performance computing resources.

Furthermore, Google Colab was instrumental in running our web application, which was developed using Streamlit. Streamlit is an open-source framework designed to create interactive web applications tailored to machine learning and data science projects. Its easy use and minimal coding requirements allowed us to develop and deploy a user-friendly interface rapidly. Streamlit's seamless integration with popular data science libraries facilitated the visualization of complex datasets and model outputs directly within the application, which was essential for iterative testing and obtaining user feedback.

Deploying a Streamlit app on Google Colab was achieved with minimal configuration, leveraging port forwarding techniques and public URLs. This setup combined the computational power of Colab with the accessibility of Streamlit's front end, resulting in an efficient and user-friendly platform. Users could interact with our models and visualizations directly from their web browsers without local installations or downloads.

This approach not only optimized our productivity but also significantly improved the accessibility and usability of our project. By leveraging cloud-based resources and streamlined application development tools, we have made advanced modeling accessible to a broader audience, ensuring that anyone with internet access can easily share and utilize our work.

As shown in Fig. 14, it depicts the user interface of the FITMI web application, showcasing the critical elements involved in the virtual try-on process. The interface prominently features sections where users can input various

Table 3 Quantitative results per category on the Dress-Code dataset [23]

Model	Upper-body		Lower-body		Dresses	
	FID u ↓	KID u ↓	FID u ↓	KID u ↓	FID u ↓	KID u ↓
PF-AFN [10]	14.32	–	18.32	–	13.59	–
HR-VTON [16]	16.86	–	22.81	–	16.12	–
CP-VTON [29]	48.31	35.25	51.29	38.48	25.94	15.81
FITMI	13.26	2.67	14.80	3.13	13.40	2.50

Fig. 10 Qualitative results generated by FITMI on Dress-Code [23] and Viton-HD [5] datasets

Fig. 11 FITMI's recommendation system for upper-body garments

details to personalize their experience. The interface features fields for selecting the garment category, specifying gender, and uploading two images: the garment and the user.

After submitting these inputs, the application processes the data and displays the virtual try-on result within approximately 3 to 4 min. Below the output, users find the recommendation systems, which provide personalized

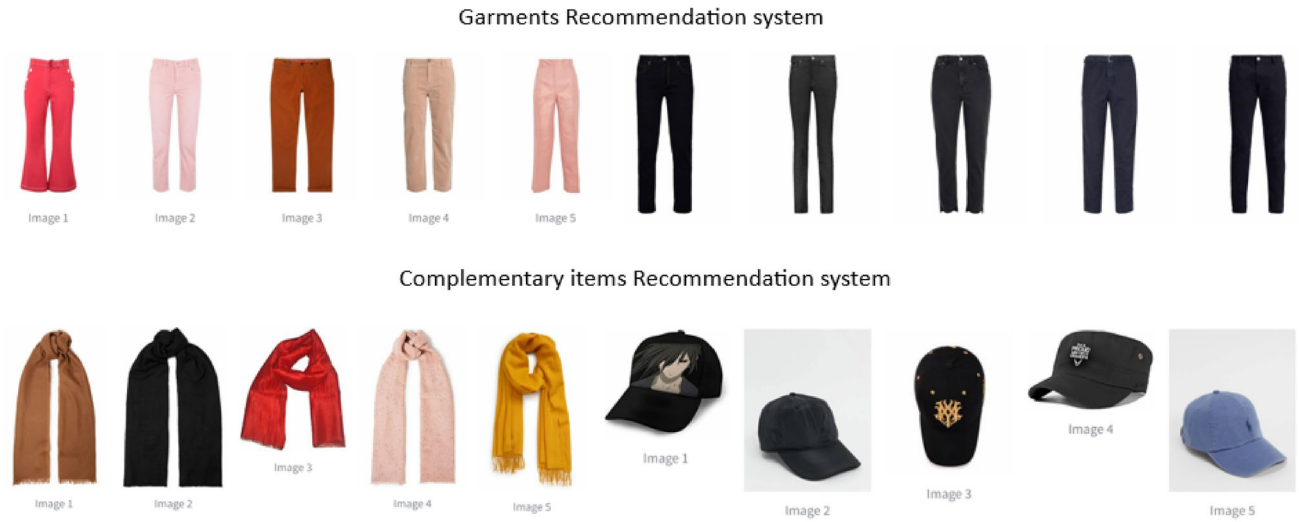

Fig. 12 FITMI's recommendation system for lower-body garments

Fig. 13 FITMI's recommendation system for dresses

suggestions based on the garment and user preferences, as shown in Fig. 15.

4.5.1 Real-time deployment feasibility

The deployment feasibility of FITMI for real-time use is crucial, particularly given the high computational demands associated with LDMs. Currently, FITMI faces challenges in optimizing performance for real-time applications due to the substantial computational intensity of LDMs, especially when processing high-resolution images or complex garment transformations. This complexity results in significant processing times that hinder real-time capabilities, particularly on consumer-grade hardware. Furthermore, the resource requirements for running LDMs typically necessitate powerful GPUs with large memory capacities, making real-time implementation on personal devices infeasible. To enhance FITMI's performance for real-time use, several optimization strategies could be employed. Implementing model compression techniques, such as

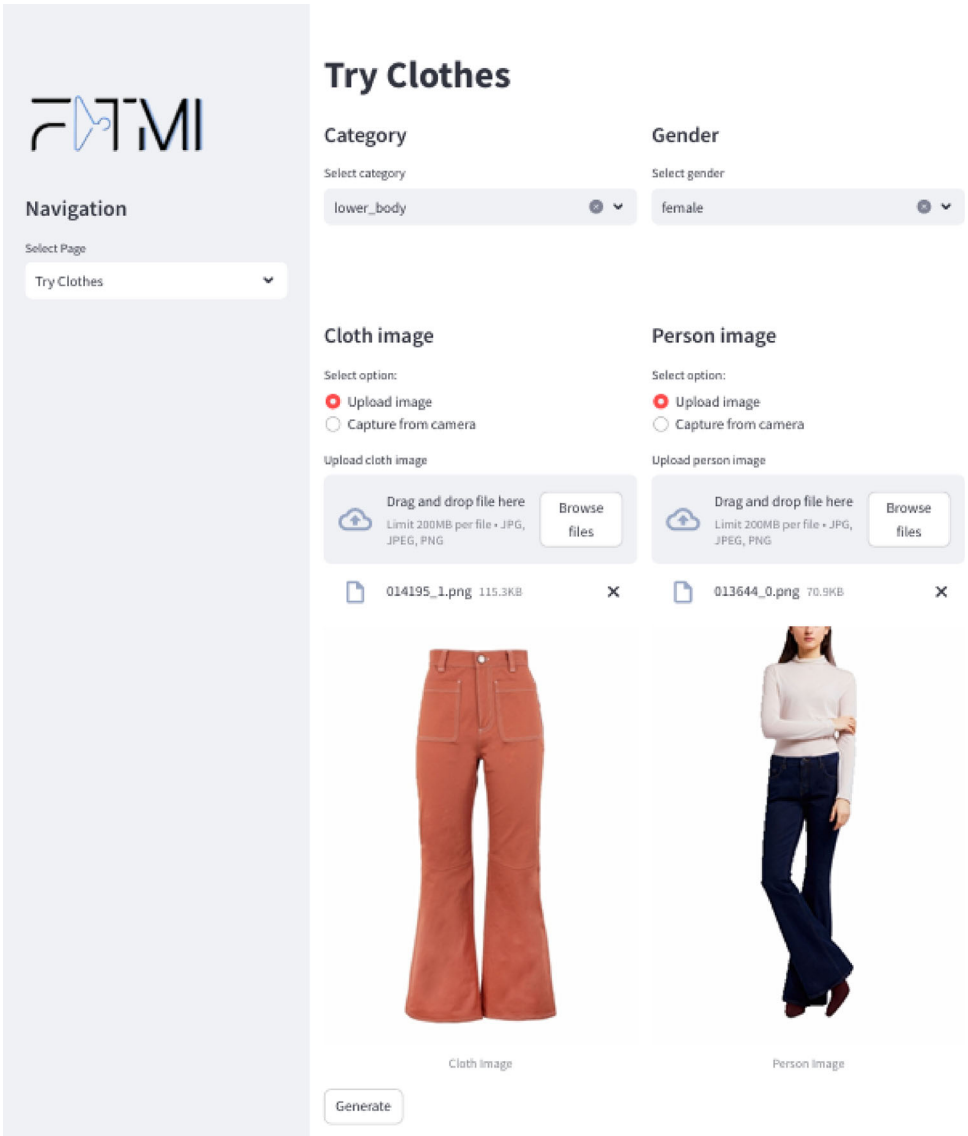

Fig. 14 The user interface of FITMI web application

pruning, quantization, and knowledge distillation, could reduce the model's size and complexity without sacrificing accuracy, thereby decreasing memory and computational requirements. Additionally, leveraging efficient inference frameworks can significantly speed up model execution on specialized hardware. A hybrid processing approach may also be beneficial, where less resource-intensive tasks, like pose estimation and garment parsing, are conducted on the user's device, while more computationally intensive operations, such as garment warping and inpainting, are offloaded to cloud servers. This balance between local processing and cloud computation could facilitate near-real-time experiences, making FITMI a more scalable and practical solution for virtual try-on applications.

5 Discussion

In developing FITMI, a well-considered approach was taken to ensure that the system not only meets the desired performance standards but also leverages the strengths of existing datasets and models. A deep understanding of the capabilities and limitations of various datasets and the specific requirements of our FITMI approach guided the choices made throughout the process.

This section comprises two aspects of our methodology: Dataset Selection Criteria and Parsing and Pose Estimation Strategy. Dataset Selection Criteria discusses the rationale behind choosing the Viton-HD [5] and Dress-Code [23] datasets, highlighting how their complementary strengths were harnessed to achieve more accurate and diverse garment simulations. The Parsing and Pose Estimation Strategy explores the decision-making process regarding the

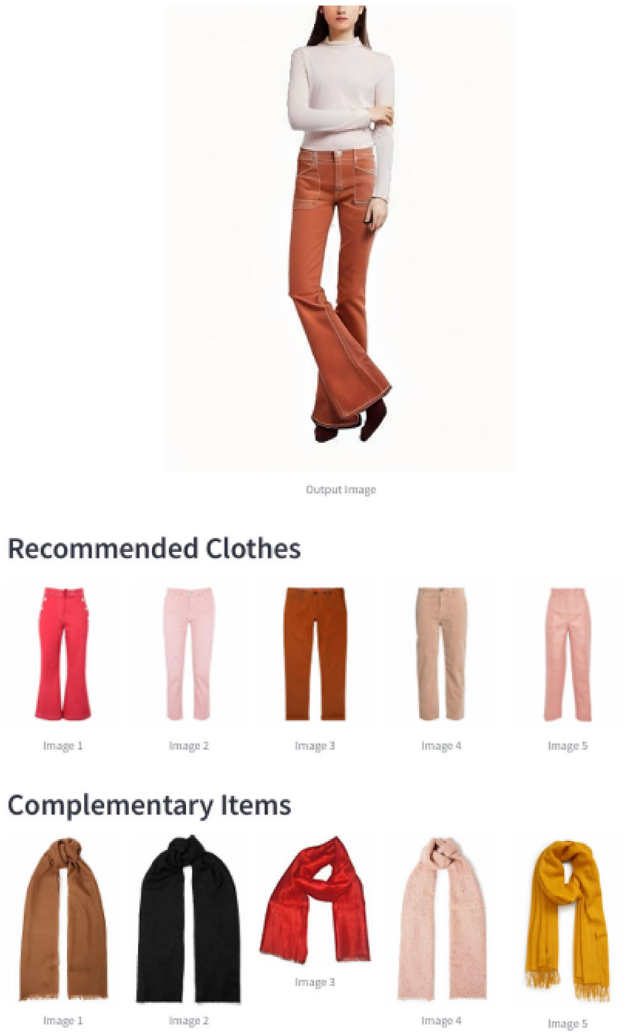

Fig. 15 The user interface of FITMI web application

parsing models and pose estimation parameters used for each dataset. This includes a detailed explanation of how the selected models and techniques were tailored to optimize the performance of FITMI across different garment categories.

5.1 Dataset selection criteria

While developing FITMI, the decision to work with both the Viton-HD [5] and Dress-Code [23] datasets was driven by each dataset's distinct strengths and limitations. Viton-HD [5] is tailored explicitly for upper-body garments, providing highly accurate and realistic results in this category. However, it does not cover lower-body clothes or dresses, which would have limited the scope of our application. On the other hand, Dress-Code [23] includes three categories—upper-body, lower-body, and dresses—making it a more versatile dataset. Nevertheless, our preliminary experiments revealed that Dress-Code [23] delivers less accurate results for upper-body garments than Viton-HD [5]. To ensure the highest quality and accuracy, Viton-HD [5] was used for upper-body garments and Dress-Code [23] for lower-body garments and dresses. This combination allowed us to leverage the strengths of each dataset.

5.1.1 Limitations in terms of fabric complexity

However, garments with intricate patterns or textures present additional challenges for FITMI, particularly in maintaining the integrity of fabric details during the warping process. Complex patterns may stretch or distort unnaturally when adjusted to fit the user's body, detracting from the garment's realistic appearance. Furthermore, the system can struggle with inconsistent texture mapping, especially when significant warping is required, leading to a failure in replicating the intended details of the fabric. Additionally, the quality of input images significantly impacts FITMI's performance, especially under challenging lighting conditions. Harsh lighting and shadows can obscure key body features, resulting in inaccuracies in pose estimation and garment alignment. Moreover, variations in lighting may alter the color representation of the garment, causing discrepancies between its appearance in the virtual environment and its actual look. Together, these factors highlight the complexities FITMI faces in delivering a truly realistic virtual try-on experience.

5.2 Parsing and Pose Estimation Strategy

In selecting the appropriate parsing and pose estimation models for FITMI, the strengths and specificities of the available datasets—Dress-Code [23] and Viton-HD [5], were carefully considered. For the Dress-Code dataset [23], the Self Correction for Human Parsing model was utilized, which is pre-trained on the ATR dataset. The ATR dataset boasts 18 semantic category labels, including crucial fashion-related elements such as 'face, sunglasses, hat, scarf, hair, upper clothes, left and right arms, belt, pants, left and right legs, skirt, left and right shoes, bag, dress, and background.' This comprehensive set of labels allows for detailed and accurate parsing across the diverse garment types present in the Dress-Code dataset.

In contrast, the Viton-HD [5] parsing model is pre-trained on the LIP dataset, which provides 20 semantic category labels. In addition to the categories present in ATR, LIP includes 'Coat' and 'Jumpsuits,' making it particularly suitable for upper-body garment parsing. This model was selected for Viton-HD [5] to ensure the parsing accuracy is maximized for the categories where this dataset excels.

A key difference in our approach lies in using Dense Pose detectron2 parameters. For Viton-HD [5], the

dp_segm parameter was employed to generate precise segmentation masks for annotated persons, significantly enhancing pose estimation accuracy. This parameter is especially effective in maintaining high fidelity in upper-body garment simulations. On the other hand, for the Dress-Code dataset [23], the dp_u parameter was employed, facilitating point annotation colored according to their U coordinate in part parameterization. This approach is crucial for accurately representing lower-body garments and dresses.

5.3 Addressing body diversity and complex poses

FITMI effectively addresses the challenges posed by diverse body types and complex poses through a comprehensive integration of pose mapping and garment warping techniques. The system employs advanced models, Open Pose [4] and Dense Pose [31], which are integral parts of the preprocessing pipeline. These models extract detailed pose maps that identify key body landmarks, such as joints and limbs, ensuring an accurate representation of the user's anatomy, regardless of how dynamic or complex their pose may be. A crucial component of FITMI is the geometric matching module utilized during the garment fitting process. This module computes a correlation map between the user's body and the in-shop garment. By identifying how the garment should be adjusted to align with the user's unique body shape and pose, it employs a Thin-Plate Spline (TPS) transformation [8]. This flexible geometric transformation technique allows for smooth and continuous warping of the garment to match the user's contours, effectively capturing the nuances of different body shapes and accommodating subtle details like body curvature and specific postures.

To further refine the garment fit, FITMI incorporates a U-Net-based refinement step, also utilized in the garment fitting process. This stage takes input from the coarse warped garment, the pose map, and a masked model image—an isolated version of the user's body without the original garment. The U-Net model fine-tunes the fit by adjusting details such as fabric stretching and garment alignment based on the specific pose. This ensures that the garment drapes naturally over the body, even in complex or contorted positions. Handling complex poses is a significant challenge in virtual try-on systems, but FITMI addresses this by employing pose maps with multiple key points, capturing the nuances of body movements. Additionally, FITMI's approach adapts to different body types by utilizing shape-aware garment warping techniques and modifying the garment based on unique body dimensions extracted from the pose map. Throughout the garment warping and fitting process, FITMI emphasizes preserving the user's identity, pose, and physical features, resulting in a final image that appears both realistic and natural.

6 Conclusion

In this work, we went through the detailed development and optimization of the FITMI application, leveraging the latent diffusion models (LDMs) with two well-known benchmarks: Dress-Code [23] and Viton-HD [5] datasets. By enhancing a pre-trained model, that utilized the textual inversion technique to significantly increase the detail retention of in-shop garments, demonstrating its efficacy in conditioning the generation process for more accurate representations. Moreover, the Enhanced Mask-Aware Skip Connection (EMASC) modules improved the inpainting output image quality, reducing the auto-encoder compression loss of LDMs. This advancement notably enhances the perceived quality of high-frequency human body details.

In the preprocessing phase, four critical techniques were utilized. Each method was integrated and executed consecutively to ensure optimal input quality for the virtual try-on model. Central to the FITMI application's user appeal is its sophisticated recommendation system. The application allows users to see themselves in desired outfits and intelligently recommends garments based on individual style preferences. This personalization is further enhanced by FITMI's ability to suggest complementary items, effectively simulating a personalized shopping assistant. Numerous experiments were conducted to refine these processes, achieving the best possible accuracy and realism in the virtual try-on experience. These experiments were critical in helping us determine the most effective configurations and settings for our models, ultimately leading to a more reliable and user-friendly application.

Looking ahead, several enhancements are anticipated to further improve the FITMI application. A key feature under development is real-time virtual try-on, which will allow for instant visualization of how clothes fit on a moving image, making the virtual try-on experience more dynamic and interactive. Moreover, plans are underway to expand FITMI's functionality to serve both individual consumers and businesses within the apparel industry by integrating with online retail platforms to offer a tool that not only to enhance the shopping experience for customers but also to increase engagement and reduce return rates. By continuously advancing FITMI, we aim to set a new standard for virtual try-on technologies, bridging the gap between digital and physical shopping experiences.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data availability Data are available on request from the authors.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ali H (2021) Real fashion dataset
2. Bińkowski M, Sutherland DJ, Arbel M, Gretton A (2018) Demystifying mmd gans. arXiv preprint arXiv:1801.01401,
3. Cao H, Tan C, Gao Z, Xu Y, Chen G, Heng P-A, Li SZ (2024) A survey on generative diffusion models. IEEE Trans Knowl Data Eng
4. Cao Z, Simon T, Wei S-E, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
5. Choi S, Park S, Lee M, Choo J (2021) Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14131–14140
6. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. Ieee
7. Detlefsen NS, Borovec J, Schock J, Jha AH, Koker T, Di Liello L, Stancl D, Quan C, Grechkin M, Falcon W (2022) Torchmetrics-measuring reproducibility in pytorch. J Open Sour Softw 7:4101
8. Duchon J (1977) Splines minimizing rotation-invariant semi-norms in sobolev spaces. In: Constructive Theory of Functions of Several Variables: proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976, pp 85–100. Springer
9. Gatis D (2020) Rembg: a tool to remove images background
10. Ge Y, Song Y, Zhang R, Ge C, Liu W, & Luo P (2021) Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8485–8493
11. Han X, Wu Z, Wu Z, Yu R, Davis LS (2018) Viton: an image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7543–7552
12. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv Neural Inform Process Syst 30
13. Islam T, Miron A, Liu X, Li Y (2024) Deep learning in virtual try-on: a comprehensive survey. IEEE Access
14. Khatir A, Capozucca R, Khatir S, Magagnini E, Benaissa B, Le Thanh C, Wahab MA (2023) A new hybrid pso-yuki for double cracks identification in cfrp cantilever beam. Compos Struct 311:116803
15. Khatir A, Capozucca R, Magagnini E, Oulad Brahim A, Osmani A, Khatir S, Abualigah L (2024) Advancing structural integrity prediction with optimized neural network and vibration analysis. J Struct Integr Main 9:2390258
16. Lee S, Gu G, Park S, Choi S, Choo J (2022) High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: European Conference on Computer Vision, pp 204–219. Springer
17. Li P, Xu Y, Wei Y, Yang Y (2020) Self-correction for human parsing. IEEE Trans Pattern Anal Mach Intell 44:3260–3271
18. Lin C, Li Z, Zhou S, Hu S, Zhang J, Luo L, Zhang J, Huang L, He Y (2022) Rmgn: a regional mask guided network for parser-free virtual try-on. arXiv preprint arXiv:2204.11258,
19. Lin H, Wen H, Song X, Liu M, Hu Y, Nie L (2024) Fine-grained textual inversion network for zero-shot composed image retrieval. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 240–250
20. Liu R, Balakrishnan B, Saari EM (2024) How ar technology is changing consumer shopping habits: from traditional retail to virtual fitting. Acad J Sci Technol 9:140–144
21. Morelli D, Baldrati A, Cartella G, Cornia M, Bertini M, Cucchiara R (2023a) Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 8580–8589
22. Morelli D, Baldrati A, Cartella G, Cornia M, Bertini M, Cucchiara R (2023b) Ladi-vton: latent diffusion textual-inversion enhanced virtual try-on. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 8580–8589
23. Morelli D, Fincato M, Cornia M, Landi F, Cesari F, Cucchiara R (2022) Dress code: high-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2231–2235
24. Park T, Liu M-Y, Wang T-C, Zhu J-Y (2019) Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2337–2346
25. Parmar G, Zhang R, Zhu J-Y (2022) On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11410–11420
26. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022a) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10684–10695
27. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022b) High-resolution image synthesis with latent diffusion models. In:

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10684–10695
28. Tian X, Jiang H, Zhao X (2024) Product assortment and online sales in community group-buying channel: evidence from a convenience store chain. J Retail Consum Serv 79:103838
29. Wang B, Zheng H, Liang X, Chen Y, Lin L, Yang M (2018) Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV), pp 589–604
30. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13:600–612
31. Wu Y, Kirillov A, Massa F, Lo W-Y, Girshick R (2019) Detectron2. https://github.com/facebookresearch/detectron2
32. Yang H, Zhang R, Guo X, Liu W, Zuo W, Luo P (2020) Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7850–7859
33. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 586–595