International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Multimodal AI-Powered Fashion Analysis and Recommendation System

Ms. Yasmin R. Shaikh 1 Ms. Siddhi Charudtta Sonawnae 2 Ms. Apurva Pawar 3 Ms. Payal Aanasaheb Dethe 4 Ms. Rutuja Sathish Bhandare
Lecturer, Department of Artificial Intelligence & Machine Learning, Mahavir Polytechnic, Nashik, Maharashtra, India 1
Student, Department of Artificial Intelligence & Machine Learning, Mahavir Polytechnic, Nashik, Maharashtra, India 2, 3, 4, 5

Abstract - The rapid growth of online fashion platforms has increased the need for intelligent and personalized recommendation systems. Traditional recommendation approaches such as collaborative filtering and content-based filtering often fail to capture complex user preferences and visual attributes effectively. This paper proposes a Multimodal AI-Based Fashion Analysis and Recommendation System that integrates both image and text inputs to improve recommendation accuracy and relevance. The system utilizes deep learning techniques to extract visual features from fashion images and semantic features from textual descriptions, combining them to generate meaningful product suggestions. Additionally, an Explainable Artificial Intelligence (XAI) component is incorporated to enhance transparency and user trust by providing reasoning behind recommendations. The system also includes a cloth segmentation module using UNet and a virtual try-on feature powered by the VITON-HD framework. The proposed approach aims to improve personalization, user engagement, and decision-making in online fashion platforms.

.

Key Words: Multimodal Learning, Fashion Recommendation System, Deep Learning, Explainable AI, Image Processing, Natural Language Processing, Virtual Try-On, CLIP, UNet, VITON-HD

1.INTRODUCTION

The rapid expansion of e-commerce platforms has significantly transformed the fashion industry, enabling customers to explore and purchase products online with convenience. As product catalogs grow larger and more diverse, users often face difficulty identifying items that match their preferences. This challenge has increased the importance of intelligent recommendation systems in online retail environments.

Traditional recommendation techniques such as collaborative filtering and content-based filtering primarily rely on user interaction history or predefined product attributes. While effective to some extent, these methods struggle to capture complex visual patterns, style combinations, and contextual meanings present in fashion products. Fashion recommendation is inherently multimodal, as decisions are influenced not only by textual descriptions but also by visual appearance, color combinations, patterns, and design elements.

Recent advancements in Artificial Intelligence (AI), particularly in deep learning, have enabled the integration of multiple data modalities. Multimodal learning combines visual and textual information to generate more accurate and meaningful recommendations. By leveraging convolutional neural networks (CNNs) for image feature extraction and natural language processing (NLP) techniques for textual understanding, recommendation systems can better model user preferences and product similarities.

In addition to accuracy, transparency has become a crucial requirement in modern AI systems. Users often hesitate to trust automated recommendations when the reasoning behind them is unclear. Explainable Artificial Intelligence (XAI) addresses this by providing interpretable insights into how recommendations are generated. This paper proposes a system that integrates image and text inputs along with explainability, virtual try-on, and cloth segmentation into one unified platform.

II. LITERATURE SURVEY

Radford et al. introduced CLIP (Contrastive Language-Image Pretraining) at OpenAI, a model trained on 400 million image-text pairs using a contrastive learning objective. CLIP encodes both images and text into a

shared 512-dimensional vector space, enabling semantic matching across modalities. This work forms the core of the proposed recommendation engine, as it allows the system to understand user intent expressed in natural language alongside visual preference expressed through an uploaded image.

Choi et al. proposed VITON-HD, a high-resolution virtual try-on framework that addresses the challenge of misalignment between warped clothing and the target person. VITON-HD introduces the ALIAS (ALIgnment-Aware Segment) normalization technique and operates in three stages: segmentation generation, geometric matching for cloth warping, and final image synthesis at 768×1024 resolution. This framework is directly integrated into the proposed system for the virtual try-on feature.

Ronneberger et al. introduced U-Net, an encoder-decoder convolutional neural network architecture originally designed for biomedical image segmentation. The symmetric structure with skip connections preserves spatial detail through downsampling, making it highly effective for precise boundary detection. In the proposed system, UNet is used to segment clothing items from background in uploaded images, forming the foundation of the segmentation and batch processing modules.

He et al. proposed ResNet, introducing residual connections that allow very deep neural networks to be trained without degradation. This architecture has been widely adopted as a backbone in visual feature extraction tasks and is referenced in the context of encoder architectures used within the VITON-HD pipeline. ResNet-based encoders enable robust feature representation for both the person and cloth images during try-on synthesis.

III. PROBLEM STATEMENT

In online fashion retail, customers face a fundamental limitation: they cannot physically interact with clothing before purchasing. Current e-commerce platforms provide only text search or basic image search, which are insufficient for capturing a customer's complete intent. A user may want to find a clothing item that looks similar to one they already own but is also appropriate for a specific occasion — a need that neither text search nor image search alone can satisfy.

Existing virtual try-on tools are either available only within closed commercial platforms or require specialized hardware and preprocessing pipelines that are inaccessible to general users. Recommendation systems on major platforms such as Amazon and Myntra do not provide explanations for why items are suggested, reducing user trust and making it harder for customers to make informed decisions.

There is currently no open, unified system that combines multimodal search, virtual try-on, explainable recommendations, and clothing segmentation into a single accessible application. This gap results in poor shopping experiences, higher return rates, and reduced customer satisfaction.

IV. EXISTING SYSTEM

Most existing fashion recommendation systems operate in isolation, addressing only one aspect of the problem. Text-based search systems such as those used by Amazon and Myntra allow users to describe what they want in words, but cannot interpret visual preferences. Reverse image search tools like Google Lens accept image input but cannot incorporate additional context such as occasion or style preference expressed in words.

Virtual try-on tools such as Snap's augmented reality features are available only within specific mobile applications and do not integrate with recommendation or search functionality. They also require complex preprocessing of person images including pose estimation and body parsing, making them inaccessible as standalone tools.

Existing multimodal research models such as FashionCLIP address image-text retrieval but do not provide explainable outputs, try-on capability, or a deployable user interface. These models remain as research prototypes and are not accessible to end users without significant technical expertise.

Table I: Comparison with Existing Systems

System	Text Search	Image Search	Try-On	Multimodal	Explainable AI
Amazon / Myntra	✓	✓	✗	✗	✗
Google Lens	✗	✓	✗	✗	✗
Snap AR Try-On	✗	✗	✓	✗	✗
FashionCLIP	✗	✓	✗	✓	✗
Proposed System	✓	✓	✓	✓	✓

V. PROPOSED SYSTEM

The proposed system is a unified web application built with Streamlit and PyTorch that brings together four intelligent capabilities under one interface. Figure 1 shows the complete system workflow.

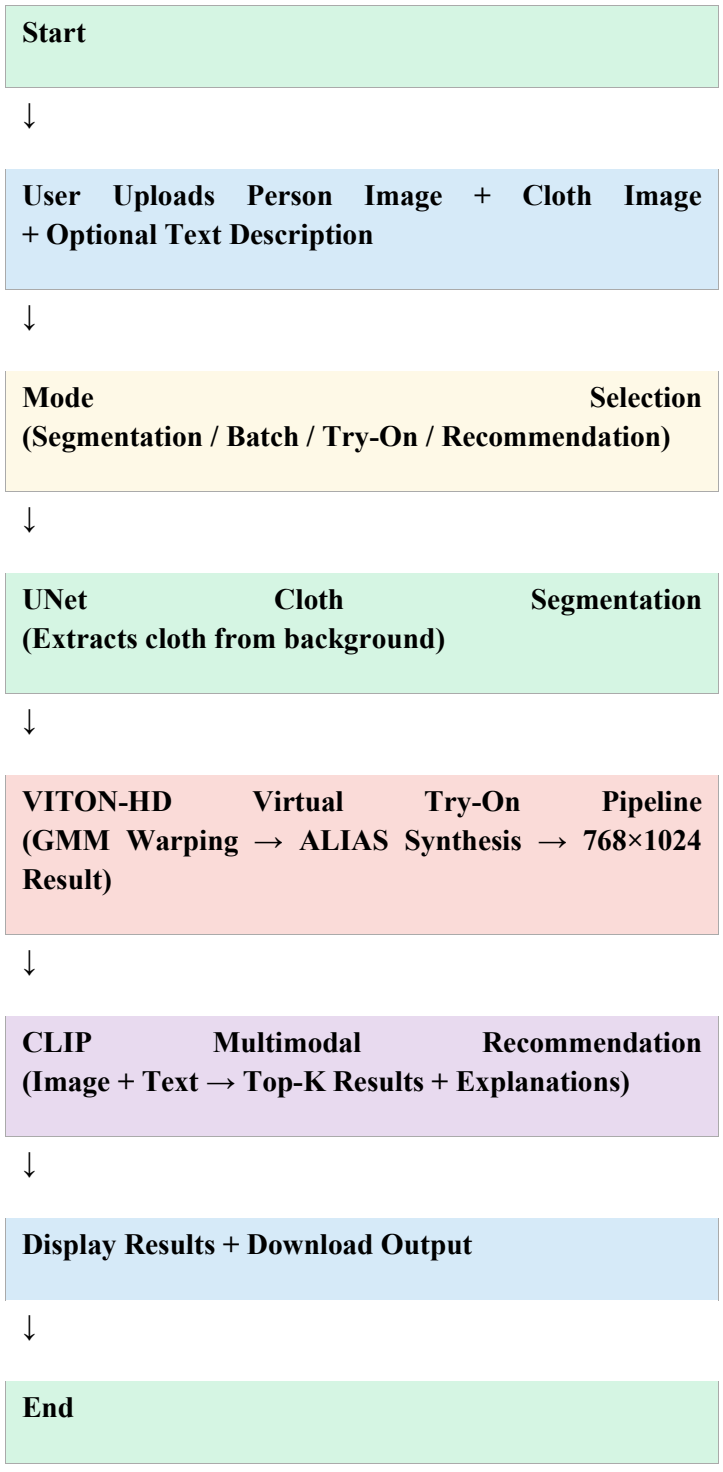

Fig. 1: Proposed System Workflow

A. Cloth Segmentation

The user uploads a clothing image. The system passes it through a trained UNet convolutional neural network which generates a binary mask separating the garment from the background. The segmented cloth image is displayed alongside statistics such as cloth area percentage and pixel count. The user can download the mask or extracted cloth image.

B. Batch Processing

Users upload multiple clothing images simultaneously. The system processes each image through the UNet segmentation pipeline and displays a summary table with results for all images. This feature is designed for catalog-scale operations where many items need to be processed at once.

C. Virtual Try-On

The user selects or uploads a person image and a cloth image. The system runs the VITON-HD three-phase pipeline: Phase 1 generates a body segmentation map, Phase 2 uses the Geometric Matching Module (GMM) to warp the cloth to the person's body shape, and Phase 3 uses the ALIAS generator to synthesize the final photorealistic try-on result at 768×1024 resolution.

D. Multimodal Recommendation

The user uploads a reference clothing image and optionally provides a text description such as "formal black dress for evening party." The system encodes both inputs using OpenAI CLIP and computes cosine similarity against all catalog items. The final recommendation score combines image similarity and text similarity with equal 0.5 weighting. Top-5 or top-10 results are displayed with explanations for each recommendation, implementing Explainable AI.

VI. OUTPUT

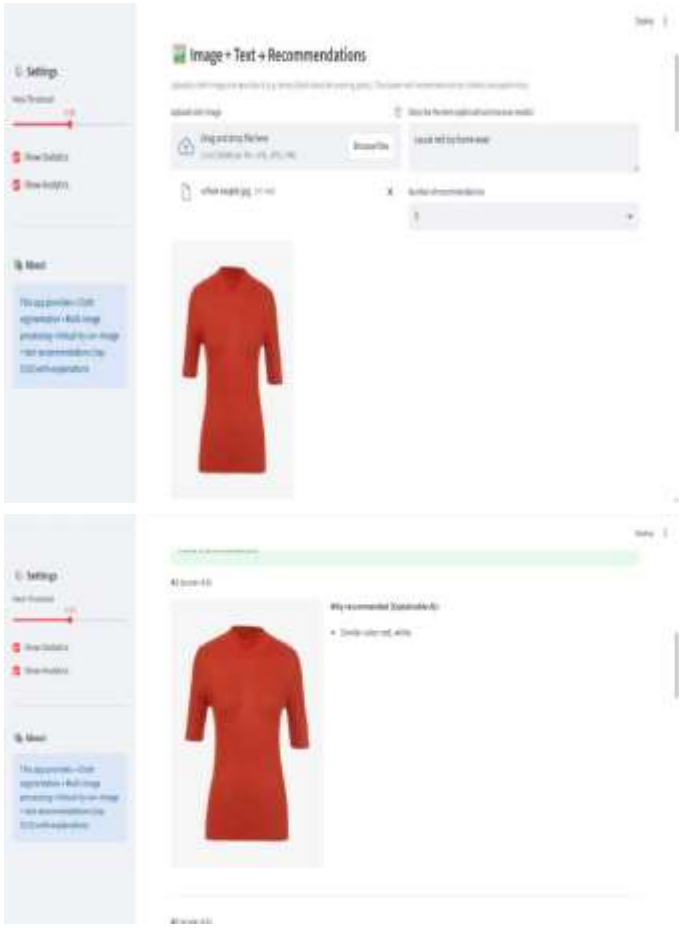

Fig. 2: Multimodal Recommendation — Query Image, Text

Fig. 3: Virtual Try-On — Person, Cloth, and Try-On Result

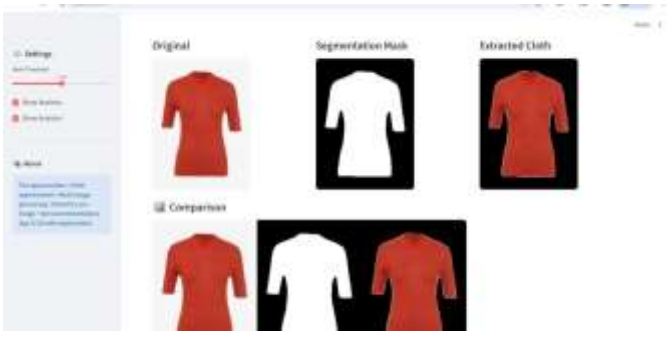

Fig. 4: Cloth Segmentation — Original, Mask, and Extracted Cloth

3. CONCLUSIONS

This paper presented a Multimodal AI-Powered Fashion Analysis and Recommendation System that integrates cloth segmentation, batch processing, virtual try-on, and CLIP-based multimodal recommendation into a single unified web application. The system addresses real limitations in online fashion retail by enabling visual try-on and semantic recommendation that goes beyond what text search or image search alone can provide.

The use of OpenAI CLIP for joint image-text encoding allows the system to understand user intent in a genuinely multimodal way, delivering more accurate and context-aware results than conventional recommendation approaches. The Explainable AI component builds user trust by providing natural language reasoning for every recommendation. The system is fully deployable on standard hardware without GPU requirements.

In future work, we plan to integrate real-time pose estimation for user-uploaded images to improve try-on quality, expand the clothing catalog with a larger annotated dataset, and develop a mobile-compatible interface with on-device CLIP inference for wider accessibility.

.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our project guide for their constant guidance, support, and encouragement throughout the development of this project. Their technical insights and suggestions greatly enhanced the quality and functionality of the system. We are also thankful to our institution for providing the necessary resources and infrastructure. Finally, we extend our appreciation to our peers and all individuals who tested the system and provided valuable feedback, which helped improve the usability and performance of the application.

REFERENCES

1. A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. ICML, 2021.
2. S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization," in Proc. CVPR, 2021.
3. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, 2015.
4. X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An Image-Based Virtual Try-On Network," in Proc. CVPR, 2018.
5. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. CVPR, 2016.
6. Z. Guo et al., "FashionCLIP: Leveraging CLIP for Fashion Retrieval," in Proc. ACM MM, 2023.
7. B. Wang et al., "Toward Characteristic-Preserving Image-based Virtual Try-On Network," in Proc. ECCV, 2018.
8. J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in Proc. CVPR, 2015.