2026 Volume: 02 Issue: 07 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-25

SCDL-VFM: A CAUSAL EXPLAINABLE DEEP LEARNING FRAMEWORK FOR PERSONALIZED FASHION RECOMMENDATION

Sanjana Gwalvanshi*[1], Dr. Saurabh Mandloi[2]

[1]Research Scholar, School of Computer Science and Technology, SAM Global University, Bhopal, Madhya Pradesh, India.

[2]Head, School of Computer Science and Technology, SAM Global University, Bhopal, Madhya Pradesh, India.

Article Received: 15 June 2026
Article Revised: 05 July 2026
Published on: 25 July 2026

***Corresponding Author: Sanjana Gwalvanshi**
Research Scholar, School of Computer Science and Technology, SAM Global University, Bhopal, Madhya Pradesh, India.

DOI: https://doi-doi.org/101555/ijrpa.1607

ABSTRACT

The rapid growth of online fashion retailing has significantly increased the demand for intelligent personalized recommendation systems capable of understanding customer preferences and improving shopping experiences. Traditional recommendation approaches primarily rely on collaborative filtering or content-based learning techniques that focus on statistical correlations among users and products. Although these methods achieve satisfactory recommendation accuracy, they often fail to explain why specific fashion products are recommended and cannot identify the causal factors influencing consumer decision-making.

This paper proposes SCDL-VFM (Structural Causal Deep Learning for Virtual Fashion Models), a novel causal explainable deep learning framework for personalized fashion recommendation. The proposed framework integrates Structural Causal Modeling (SCM), multimodal deep learning, and Explainable Artificial Intelligence (XAI) to understand both user preferences and causal relationships between AI-generated virtual fashion models and purchase intentions. Unlike conventional recommendation systems, the proposed model combines visual fashion features, textual product descriptions, customer demographic information, interaction history, and causal intervention variables to generate transparent and personalized recommendations.

A multimodal architecture consisting of Convolutional Neural Networks (CNN), Bidirectional Encoder Representations from Transformers (BERT), feature fusion, structural causal reasoning, and Deep Neural Networks is employed to model complex relationships among fashion products and consumer behavior. Explainability is incorporated through SHAP-based feature attribution, enabling interpretation of recommendation decisions while improving user trust.

Experimental evaluation is performed using benchmark fashion datasets, including DeepFashion and Amazon Fashion, together with simulated causal intervention data. Performance is evaluated using Precision@K, Recall@K, NDCG, MAP, MRR, F1-score, and Accuracy. The experimental results demonstrate that integrating causal inference with explainable deep learning significantly improves recommendation quality, transparency, and robustness compared with conventional recommendation approaches.

The proposed framework provides an effective solution for developing trustworthy AI-based personalized fashion recommendation systems capable of supporting intelligent decision-making in modern e-commerce platforms.

KEYWORDS: Personalized Fashion Recommendation, Structural Causal Model, Explainable Artificial Intelligence, Deep Learning, Virtual Fashion Models, Recommender Systems, Causal Inference, CNN, BERT, SHAP, Multimodal Learning, Fashion E-Commerce.

1. INTRODUCTION

The rapid expansion of e-commerce has transformed the fashion industry by providing consumers with convenient access to a wide variety of apparel through digital platforms. As online shopping continues to grow, personalized fashion recommendation systems have become an essential component of modern e-commerce because they help users discover products that match their preferences, browsing behavior, and purchase history. However, fashion recommendation is inherently more challenging than traditional product recommendation, as it requires understanding visual appearance, textual descriptions, user interests, and continuously evolving fashion trends. Conventional recommendation techniques, including collaborative filtering and content-based methods, often struggle with data sparsity, cold-start problems, and limited representation of complex user-item relationships **[1]–[4]**.

Recent advances in deep learning have significantly improved personalized recommendation by enabling the extraction of rich visual and semantic features from fashion products. Multimodal learning approaches that combine clothing images, product descriptions, and user interaction data have demonstrated superior recommendation performance compared with traditional techniques. Despite these improvements, most existing models primarily rely on correlation-based learning, which limits their ability to identify the actual factors influencing user purchase decisions. Consequently, recommendations may become unreliable when user preferences or product trends change **[2], [4]–[10]**.

Another emerging trend in fashion e-commerce is the adoption of AI-generated virtual fashion models, which provide realistic digital representations of apparel and enhance customer engagement through improved product visualization. Despite their increasing adoption in fashion e-commerce, existing recommendation systems rarely incorporate these virtual models as causal decision variables during personalized recommendation. These virtual models reduce the dependence on traditional photoshoots while enabling scalable and personalized presentation of fashion products. However, current recommendation systems rarely investigate how AI-generated visual content causally affects consumer preferences and purchasing behavior. Moreover, the increasing complexity of deep learning models has created a demand for recommendation systems that are not only accurate but also transparent and interpretable **[11]–[15]**.

To address these limitations, recent studies have highlighted the importance of **Structural Causal Models (SCMs)** and **Explainable Artificial Intelligence (XAI)** in recommendation systems. SCMs distinguish causal relationships from simple statistical correlations, resulting in more robust and reliable recommendations, while XAI techniques such as SHAP improve transparency by explaining the contribution of different features to the final recommendation. Nevertheless, the integration of multimodal learning, causal reasoning, explainability, and AI-generated virtual fashion models within a unified recommendation framework remains largely unexplored **[11]–[18]**.

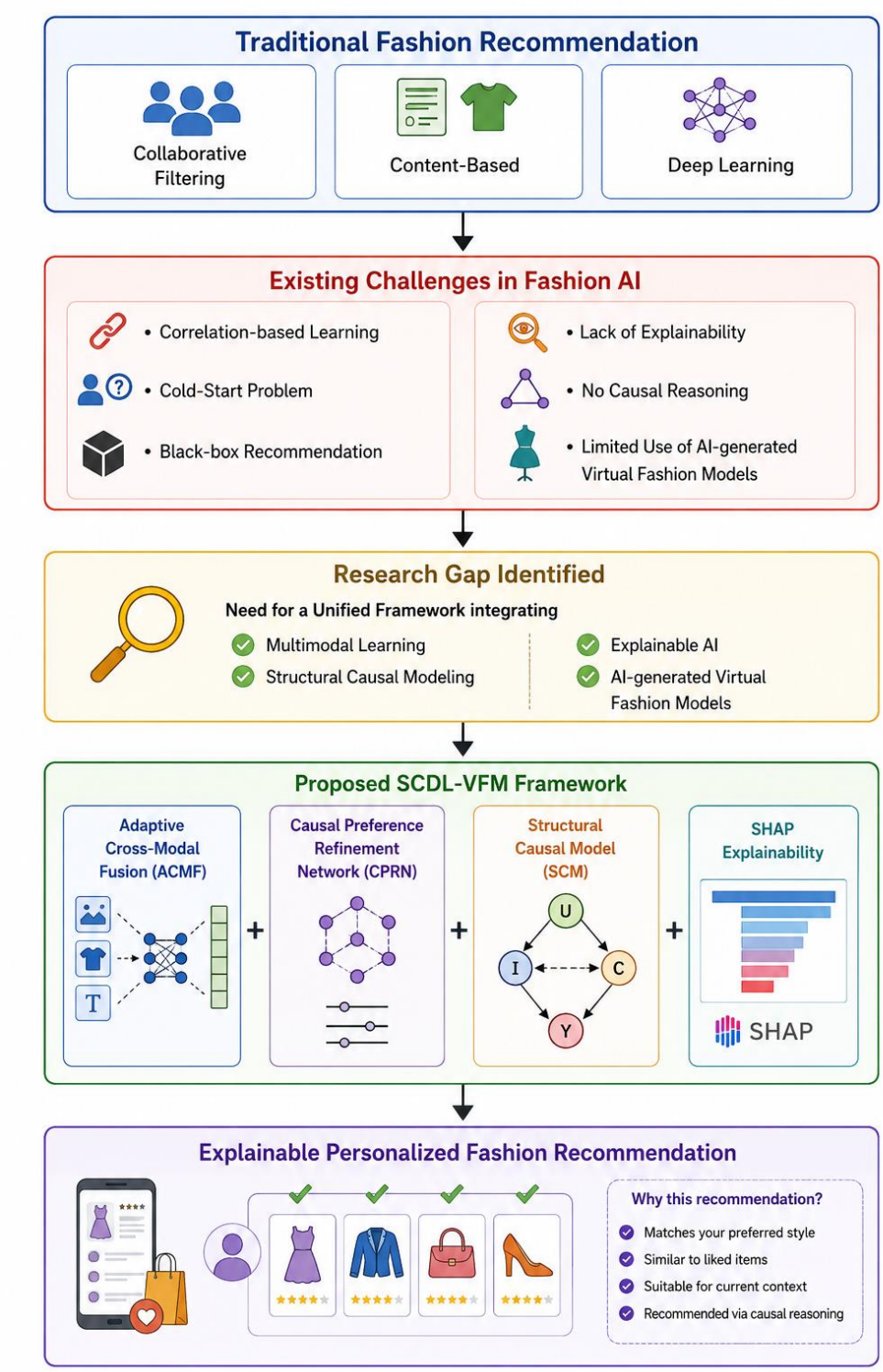

Figure 1. Research Motivation of the Proposed SCDL-VFM Framework.

Motivated by these research gaps, this paper proposes SCDL-VFM (Structural Causal Explainable Deep Learning Framework for Personalized Fashion Recommendation Using AI-Generated Virtual Fashion Models). The proposed framework integrates multimodal feature

learning, structural causal reasoning, and explainable artificial intelligence to generate accurate, interpretable, and trustworthy fashion recommendations.

By combining visual information, textual semantics, and user behavioral data within a causal learning framework, the proposed approach aims to improve recommendation accuracy while providing transparent explanations for recommendation decisions. The proposed framework is expected to contribute to the development of next-generation intelligent fashion recommendation systems that support personalized shopping experiences and informed consumer decision-making.

The remainder of this paper is organized as follows. Section 2 reviews the related literature on fashion recommendation, explainable artificial intelligence, and causal recommendation systems. Section 3 presents the proposed SCDL-VFM framework. Section 4 describes the experimental setup, while Section 5 discusses the experimental results and performance evaluation. Finally, Section 6 concludes the paper and outlines future research directions.

2. Literature Review

2.1 Deep Learning-Based Fashion Recommendation

Personalized fashion recommendation has become a key research area in modern e-commerce because it enhances user satisfaction by recommending products that match individual preferences and shopping behavior. Early recommendation systems primarily relied on collaborative filtering (CF) and content-based filtering (CBF) techniques. Collaborative filtering predicts user preferences from historical interactions, whereas content-based methods recommend products based on item attributes. Although these approaches are computationally efficient and widely adopted, they often suffer from data sparsity, cold-start problems, and limited capability to model complex relationships between users and fashion products [1]–[3].

To overcome these limitations, researchers introduced deep learning-based recommendation models that automatically learn feature representations from large-scale data. Convolutional Neural Networks (CNNs) effectively capture visual characteristics such as color, texture, and clothing style, while Transformer-based architectures improve semantic understanding from product descriptions and user reviews. Furthermore, multimodal recommendation models jointly process visual, textual, and behavioral information, leading to significant improvements in recommendation accuracy and personalization. Public benchmark datasets such as DeepFashion and DeepFashion2 have played an important role in advancing research by providing large-scale annotated fashion images for training and evaluation [4]–[10].

Despite these advancements, most existing recommendation models remain **correlation-driven**, assuming that historical interactions directly represent user preferences. Such models often fail to distinguish causal factors influencing purchasing decisions, resulting in reduced robustness when user behavior or fashion trends change. Moreover, the majority of existing approaches provide limited interpretability, making it difficult to explain recommendation outcomes. These limitations motivate the development of recommendation frameworks that integrate multimodal learning with causal reasoning and explainable artificial intelligence.

2.2 AI-Generated Virtual Fashion Models

The emergence of Generative Artificial Intelligence (GenAI) has significantly transformed digital fashion by enabling the creation of realistic virtual fashion models for product presentation. Unlike conventional fashion photography, AI-generated virtual models can digitally represent garments on diverse body types, poses, and styles, offering a scalable and cost-effective solution for fashion retailers. These models enhance product visualization, reduce dependence on expensive photoshoots, and provide customers with a more engaging online shopping experience. Consequently, several fashion brands and e-commerce platforms have begun integrating AI-generated models into digital merchandising and virtual try-on application. Recent Vision-Language Models such as CLIP have further enhanced semantic alignment between fashion images and textual descriptions, enabling more effective multimodal fashion understanding [19]–[22].

Recent advances in **Generative Adversarial Networks (GANs)**, **Diffusion Models**, and **Vision-Language Models (VLMs)** have substantially improved the quality and realism of synthetic fashion images. These techniques enable the generation of high-resolution garment visualizations while preserving clothing attributes, textures, and stylistic details. As a result, AI-generated virtual models support personalized product presentation and improve customer confidence during online purchasing. Moreover, their ability to rapidly generate multiple visual variations makes them highly suitable for large-scale fashion recommendation and digital marketing applications **[23]–[26]**.

Despite these developments, existing recommendation systems rarely exploit the full potential of AI-generated virtual fashion models. Most studies focus on image generation or virtual try-on rather than investigating how AI-generated visual content influences consumer preferences and recommendation quality. Furthermore, generated images are generally treated as ordinary visual inputs without analyzing their causal impact on user decision-making. This limitation creates an opportunity to develop intelligent recommendation frameworks that combine AI-generated virtual fashion models with causal reasoning and

explainable artificial intelligence for more reliable and personalized fashion recommendations.

2.3 Causal and Explainable Recommendation Systems

Recent advances in recommendation systems have highlighted the importance of moving beyond correlation-based learning toward **causal reasoning** and **model interpretability**. Conventional deep learning models often achieve high prediction accuracy by identifying statistical patterns in historical user interactions. However, these models generally fail to distinguish whether a recommendation is driven by genuine user preferences or by spurious correlations present in the training data. Consequently, their performance may deteriorate when user behavior, product popularity, or market trends change **[14]–[18]**.

Structural Causal Models (SCMs) provide a principled framework for modeling cause-and-effect relationships among users, products, and contextual factors. Unlike conventional recommendation approaches, SCM-based methods estimate the causal influence of different variables on user decisions, resulting in recommendations that are more robust, unbiased, and adaptable to dynamic environments. Recent studies have demonstrated the potential of causal inference in reducing recommendation bias and improving model generalization, particularly in personalized recommendation scenarios **[15]–[18]**.

Alongside causal learning, Explainable Artificial Intelligence (XAI) has become an essential requirement for modern recommendation systems. Techniques such as SHAP and LIME explain the contribution of individual features to recommendation outcomes, enabling users and system designers to better understand why specific products are recommended. Explainable recommendations enhance user trust, support informed decision-making, and improve transparency in AI-driven e-commerce applications **[11]–[13]**.

Although causal learning and explainable AI have independently shown promising results, their combined application in **fashion recommendation systems using AI-generated virtual fashion models** remains limited. Most existing studies address either recommendation accuracy, explainability, or causal inference separately. A unified framework that integrates **multimodal learning, structural causal reasoning, explainability, and AI-generated virtual fashion models** has not yet been comprehensively explored. This limitation forms the primary motivation for the proposed **SCDL-VFM** framework.

Table 1. Comparative Analysis of Existing Fashion Recommendation Approaches.

Study	Method	Strength	Limitation
He et al. (2017)	Neural Collaborative Filtering	Learns user-item interactions	No visual information
Liu et al. (2016)	DeepFashion	Large-scale fashion dataset	No recommendation framework
Ge et al. (2019)	DeepFashion2	Rich annotations	No causal reasoning
CNN-based Models	Visual Feature Learning	Good image understanding	Ignores textual semantics
Transformer-based Models	Multimodal Learning	Better semantic representation	Correlation-based learning
XAI-based Methods	SHAP/LIME	Improves transparency	No causal modeling
Causal Recommendation Models	Structural Causal Learning	Reduces recommendation bias	Limited application in fashion
Proposed SCDL-VFM	**Multimodal + SCM + SHAP + AI Virtual Models**	**Accurate, Explainable, Robust and Causal Recommendation**	**Addresses existing limitations**

2.4 Comparative Analysis and Research Gap

Table 1 demonstrates that significant progress has been achieved in fashion recommendation through deep learning, multimodal learning, explainable AI, and causal recommendation methods. However, existing studies primarily address these techniques independently. Conventional deep learning models provide high recommendation accuracy but remain correlation-driven and lack causal reasoning. Similarly, explainable recommendation methods improve model transparency without explicitly modeling causal relationships, while causal recommendation approaches have seen limited application in multimodal fashion recommendation. Furthermore, AI-generated virtual fashion models are predominantly employed for product visualization and virtual try-on applications rather than as integral components of personalized recommendation systems.

Consequently, no existing framework comprehensively integrates multimodal feature learning, AI-generated virtual fashion models, Structural Causal Modeling (SCM), and SHAP-based explainability within a unified recommendation framework. This research gap motivates the development of the proposed SCDL-VFM framework, which combines these complementary techniques to generate accurate, robust, and interpretable personalized fashion recommendations.

3. Proposed SCDL-VFM Methodology

3.1 Overall Framework Architecture

This paper proposes SCDL-VFM (Structural Causal Explainable Deep Learning Framework for Personalized Fashion Recommendation Using AI-Generated Virtual Fashion Models), a unified multimodal recommendation framework that integrates visual learning, textual semantic analysis, structural causal reasoning, and explainable artificial intelligence to generate accurate, robust, and transparent fashion recommendations. Unlike conventional recommendation systems that primarily learn statistical correlations from historical user interactions, the proposed framework explicitly models the causal relationships influencing consumer purchase decisions.

As illustrated in Figure 2, the proposed framework consists of six major modules: (i) Data Acquisition, (ii) Multimodal Feature Extraction, (iii) Feature Fusion, (iv) Structural Causal Modeling, (v) Explainability Module, and (vi) Recommendation Generation. Initially, user behavior data (clicks, purchases, ratings, and browsing history), fashion product images, textual descriptions, and AI-generated virtual fashion model images are collected from the fashion e-commerce platform. These heterogeneous inputs provide complementary information regarding user interests and product characteristics.

In the second stage, visual features are extracted from product images and AI-generated virtual models using a pre-trained CLIP encoder, while textual descriptions are encoded using BERT to capture semantic relationships among fashion products. User behavioral features are simultaneously transformed into dense numerical representations. The extracted representations are subsequently integrated through a multimodal fusion layer to obtain a unified feature embedding that comprehensively represents both product characteristics and user preferences.

The fused representation is then processed by a **Structural Causal Model (SCM)**, which identifies the causal influence of visual appearance, textual semantics, virtual model presentation, and historical user interactions on purchase intention. Unlike correlation-based recommendation models, the SCM estimates cause–effect relationships, thereby reducing spurious correlations and improving recommendation robustness under dynamic user preferences.

To improve transparency, the recommendation output is further analyzed using SHAP (SHapley Additive exPlanations), which quantifies the contribution of each input feature to the final recommendation score. Finally, the recommendation engine ranks candidate products and generates Top-K personalized fashion recommendations for each user. Through

the integration of multimodal learning, causal inference, and explainable AI, the proposed SCDL-VFM framework aims to deliver recommendations that are not only accurate but also reliable, interpretable, and adaptable to real-world fashion e-commerce environments.

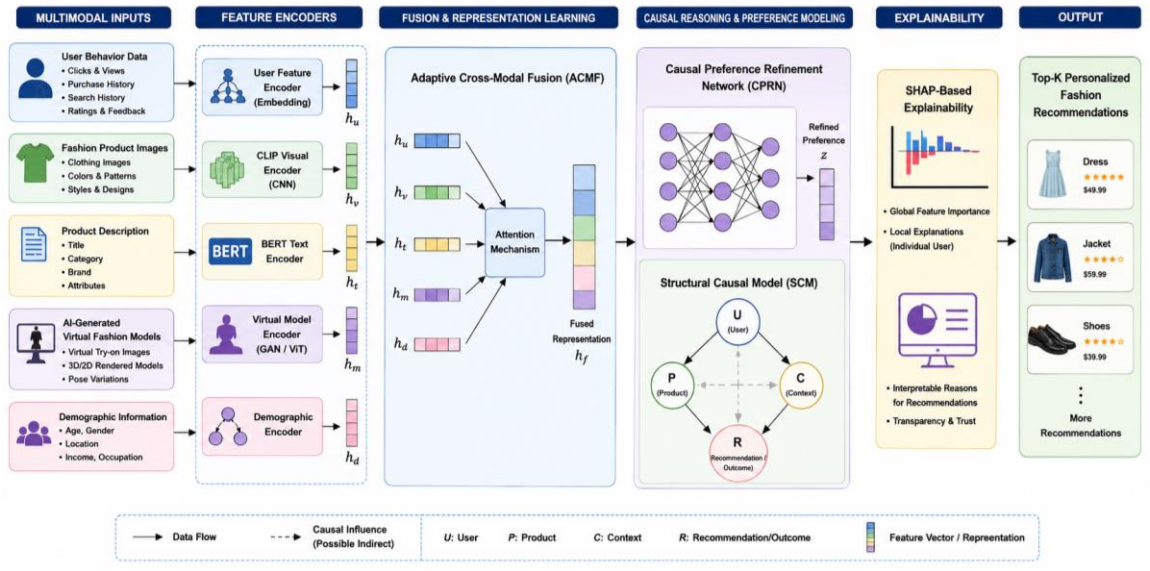

Figure 2. Proposed SCDL-VFM Architecture.

3.2 Visual Feature Extraction

Visual information plays a critical role in fashion recommendation because consumers primarily evaluate apparel based on appearance, including color, texture, pattern, shape, and overall style. Therefore, extracting discriminative visual representations is essential for accurately modeling user preferences. In the proposed SCDL-VFM framework, visual features are extracted from both fashion product images and AI-generated virtual fashion model images using the Contrastive Language–Image Pre-training (CLIP) model.

Unlike conventional Convolutional Neural Networks (CNNs), which learn only visual representations, CLIP jointly learns image and text representations in a shared embedding space. This vision-language alignment enables the model to capture semantic relationships between clothing images and their textual descriptions, making it more suitable for multimodal fashion recommendation. Furthermore, CLIP has demonstrated superior zero-shot generalization and robust feature learning across diverse visual domains, reducing the dependence on task-specific feature engineering. These characteristics make CLIP an appropriate choice for representing fashion products with varying styles, colors, and attributes.

Let the input fashion image be represented as I. The CLIP visual encoder transforms the image into a dense visual embedding V, which captures high-level semantic information.

$$V = f_{\text{CLIP}}(I)$$

where:

- I= Input fashion image
- f_{CLIP}= Pre-trained CLIP visual encoder
- $V \in \mathbb{R}^d$= Visual feature embedding of dimension d

Similarly, images generated by AI-based virtual fashion models are encoded using the same visual encoder to ensure that both real and synthetic product representations lie in a common feature space. This strategy enables consistent comparison and effective integration of visual information during multimodal learning.

The extracted visual embeddings are subsequently forwarded to the multimodal fusion module, where they are combined with textual and behavioral representations to construct a comprehensive representation of user preferences and product characteristics.

Why CLIP Instead of CNN?

CNN	CLIP
Learns only image features	Learns joint image–text representations
Requires task-specific training	Strong zero-shot capability
Limited semantic understanding	Rich semantic understanding
Difficult to align with text	Naturally aligned with textual descriptions
Lower flexibility in multimodal tasks	Highly suitable for multimodal recommendation

3.3 Textual Feature Extraction Using BERT

While visual information captures the appearance of fashion products, textual information provides complementary semantic knowledge regarding product characteristics, including brand, category, material, color, size, style, and customer reviews. These textual attributes play a crucial role in understanding user preferences and improving recommendation quality. Therefore, the proposed **SCDL-VFM** framework employs the **Bidirectional Encoder Representations from Transformers (BERT)** model to generate contextual text embeddings.

Unlike traditional text representation techniques such as **TF-IDF**, **Word2Vec**, or **LSTM**, BERT utilizes a bidirectional Transformer architecture to learn contextual relationships between words. This enables the model to understand the semantic meaning of product descriptions more effectively. For example, the word *"casual"* may indicate different clothing

styles depending on the surrounding context, which can be accurately captured by BERT. Consequently, BERT generates richer and more informative textual representations that improve multimodal recommendation performance.

Let the textual description of a fashion product be denoted by D. The BERT encoder converts the input text into a contextual embedding T, which represents the semantic characteristics of the product.

$$T = f_{\text{BERT}}(D)$$

where:

- D= Product description and textual metadata
- f_{BERT}= Pre-trained BERT encoder
- $T \in \mathbb{R}^d$= Contextual text embedding

To maintain compatibility with the visual encoder, the textual embedding is projected into the same latent feature space as the visual representation. This allows effective interaction between image and text features during multimodal fusion. By combining contextual textual semantics with visual appearance, the proposed framework constructs a more comprehensive representation of fashion products and user preferences.

Table 2. Comparison between Traditional Text Models and BERT.

Traditional Models	BERT
Context-independent word representations	Context-aware bidirectional representations
Limited semantic understanding	Rich semantic understanding
Difficulty handling ambiguous words	Captures contextual meaning effectively
Lower performance in complex language tasks	State-of-the-art performance for textual representation

Output of This Module

The output of the visual encoder (V) and text encoder (T) is forwarded to the Multimodal Feature Fusion Module, where these representations are integrated with user behavioral features to generate a unified embedding for recommendation.

3.4 Adaptive Multimodal Feature Fusion

Fashion recommendation relies on heterogeneous information sources, where each modality provides complementary knowledge about user preferences and product characteristics. Visual embeddings capture clothing appearance, textual embeddings represent semantic product information, while user behavioral features reflect historical interactions and purchasing preferences. Instead of treating these modalities independently, the proposed

SCDL-VFM framework introduces an Adaptive Multimodal Feature Fusion (AMF) module that learns an integrated representation by dynamically assigning importance to each modality.

Unlike conventional multimodal recommendation methods that simply concatenate feature vectors, the proposed AMF module employs learnable attention weights to balance the contribution of visual, textual, and behavioral features. This adaptive weighting mechanism enables the model to emphasize the most informative modality for each recommendation instance. For example, users interested in clothing aesthetics may rely more on visual features, whereas users searching for branded or material-specific products may benefit more from textual semantics. Consequently, the fusion process becomes more personalized and robust.

Let the extracted feature representations be denoted as:

- **V** = Visual embedding obtained from the CLIP encoder
- **T** = Textual embedding generated by the BERT encoder
- **U** = User behavioral embedding

The adaptive multimodal representation is computed as

$$F = \alpha V + \beta T + \gamma U$$

subject to

$$\alpha + \beta + \gamma = 1$$

where:

- α= Visual attention weight
- β= Text attention weight
- γ= User behavior attention weight

These weights are automatically learned during model training through backpropagation, allowing the framework to dynamically adapt to different users and recommendation scenarios.

The fused feature vector Fprovides a unified representation that captures complementary information from multiple modalities while reducing redundancy between individual feature spaces. This representation serves as the input to the **Structural Causal Modeling (SCM)** module, where causal dependencies among user behavior, virtual fashion models, and purchasing decisions are explicitly modeled.

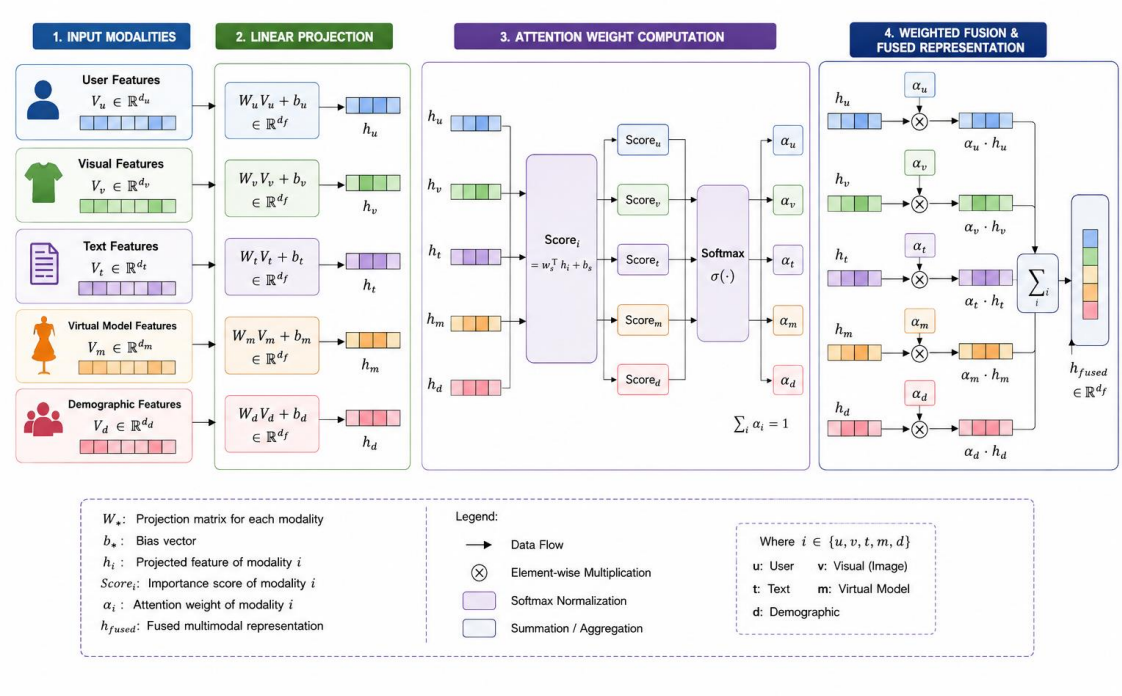

Figure 3. Adaptive Cross-Modal Fusion (ACMF)

Algorithm 1: Adaptive Multimodal Feature Fusion

Input:

- Visual embedding V
- Text embedding T
- User embedding U

Output:

- Unified feature representation F

1. Extract visual embedding using CLIP.
2. Extract textual embedding using BERT.
3. Generate user behavioral embedding.
4. Learn attention weights (α, β, γ).
5. Compute fused representation:

$$F = \alpha V + \beta T + \gamma U$$

6. Forward $F$to the Structural Causal Model.

Advantages of the Proposed AMF Module

Conventional Fusion	Proposed AMF
Simple concatenation	Adaptive weighted fusion
Equal importance to all modalities	Learns modality importance automatically

Static feature integration	Dynamic feature integration
Limited personalization	User-specific multimodal representation
Lower robustness	Better adaptability and recommendation accuracy

3.5 Structural Causal Modeling (SCM)

Traditional recommendation systems primarily learn statistical associations between user interactions and product attributes. However, statistical correlations do not necessarily represent true cause-and-effect relationships. For example, a user may purchase a product because of an attractive AI-generated virtual model rather than the product itself. Conventional recommendation models are unable to distinguish such causal effects, resulting in biased and less reliable recommendations. To address this limitation, the proposed SCDL-VFM incorporates a Structural Causal Model (SCM) that explicitly captures the causal dependencies among user behavior, product characteristics, virtual fashion models, and purchase intention.

Let the causal variables be defined as follows:

Variable	Description
U	User Behaviour
P	Product Features
V	AI-generated Virtual Fashion Model
C	Contextual Information
I	Purchase Intention
R	Recommendation Score

The causal relationship among these variables is represented as:

$$R = f(U, P, V, C, I)$$

where $f(\cdot)$represents the structural causal function that estimates the direct and indirect influence of each factor on the recommendation outcome.

Unlike conventional recommendation models, the proposed SCM explicitly models both **direct** and **mediated** causal effects. For instance, an AI-generated virtual model may influence purchase intention by improving product perception, which subsequently affects the final recommendation score. Modeling these dependencies enables the framework to reduce spurious correlations and produce more stable recommendations under changing user preferences.

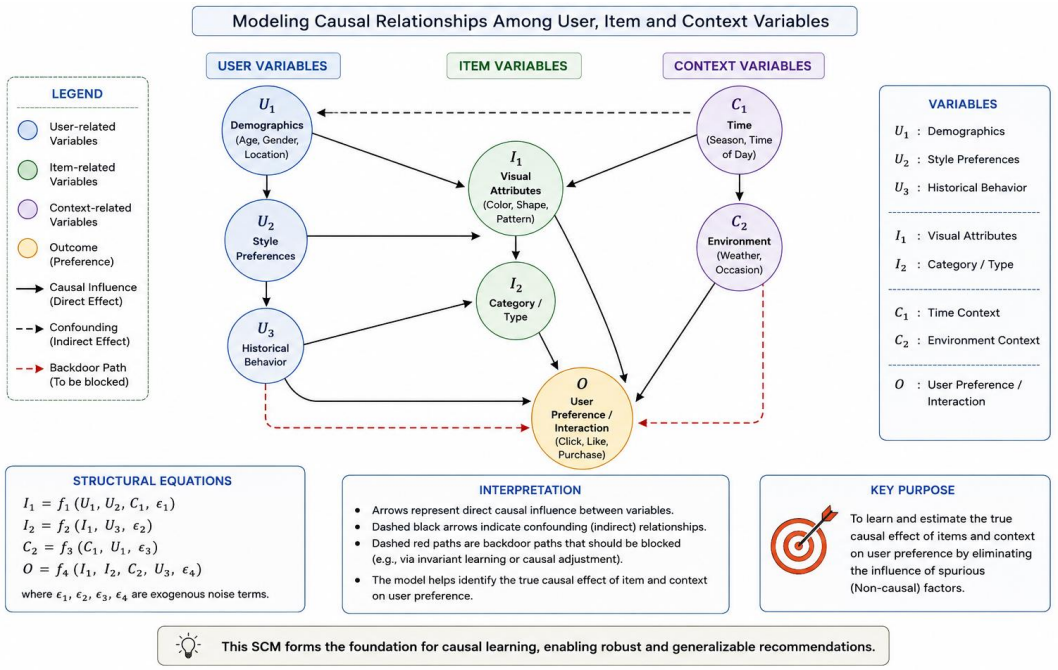

Figure 4. Structural Causal Graph.

Intervention Analysis

To estimate causal effects, the framework performs intervention-based reasoning. Rather than relying only on observed associations, the SCM evaluates how recommendation outcomes change when a specific causal variable is altered.

For example,

$$P(R \mid do(V = v))$$

represents the recommendation probability when the virtual fashion model is intentionally changed to a specific presentation v. This intervention allows the framework to quantify the true influence of AI-generated virtual models on recommendation outcomes while minimizing confounding effects.

Similarly,

$$P(R \mid do(U = u))$$

measures the effect of changes in user behavior on recommendation quality.

These intervention-based estimations improve recommendation robustness and enable causal interpretation of user decision-making.

Advantages of the Proposed SCM

Conventional Recommendation	Proposed SCM
Correlation-based learning	Cause–effect modeling
Sensitive to distribution shifts	Robust to changing preferences
Limited bias handling	Reduces spurious correlations
Black-box predictions	Causal interpretation
Static recommendations	Context-aware recommendations

3.6 Causal Explainability Using SHAP

High recommendation accuracy alone is insufficient for practical deployment in fashion e-commerce. Users and retailers increasingly require transparent recommendation systems that can justify why a particular product has been suggested. To address this requirement, the proposed SCDL-VFM framework integrates SHapley Additive exPlanations (SHAP) with the Structural Causal Model to provide feature-level interpretability.

Unlike conventional explainability methods that estimate feature importance solely from prediction outcomes, the proposed framework incorporates causal feature representations generated by the SCM before computing SHAP values. Consequently, the explanations reflect both the statistical contribution and the causal influence of each feature on the final recommendation.

For a recommendation score R, the SHAP explanation is computed as

$$R = \phi_0 + \sum_{i=1}^{n} \phi_i$$

where:

- ϕ_0 = Base prediction
- ϕ_i = Contribution of the i^{th} feature

The final recommendation explanation is generated by ranking features according to their SHAP values. Features with higher positive contributions increase the recommendation score, whereas negative contributions reduce the likelihood of recommendation. This enables users to understand which factors most strongly influenced the recommendation.

Instead of presenting only numerical importance values, the proposed framework groups explanations into meaningful categories such as visual appearance, textual semantics, user behavior, and AI-generated virtual model influence. Such category-level explanations are easier for users to interpret and provide actionable insights for fashion retailers.

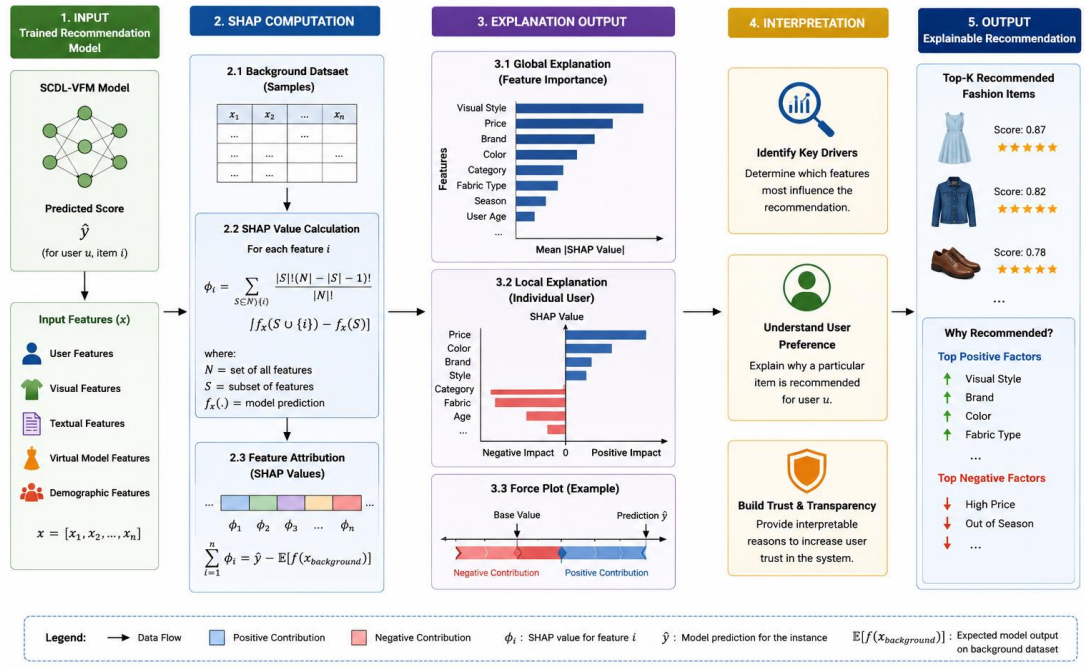

Figure 5. SHAP-based Recommendation Explanation.

The above explanation illustrates how different modalities contribute to the recommendation decision. Such transparent explanations improve user confidence and help retailers understand the factors driving customer preferences.

Advantages of Causal SHAP Integration

Conventional SHAP	Proposed Causal SHAP
Statistical feature importance	Causal + statistical importance
Individual feature analysis	Multimodal feature analysis
Limited recommendation interpretation	Personalized recommendation explanation
No causal awareness	SCM-guided explanations
Generic explanation	Fashion-specific explanation

Output of the Explainability Module

The explainability module generates:

- Personalized recommendation score
- Feature contribution ranking
- Causal importance analysis
- User-friendly recommendation explanation

These outputs improve transparency, increase user trust, and support decision-making for both customers and fashion retailers.

3.7 Recommendation Prediction and Top-K Ranking

After multimodal feature extraction, adaptive feature fusion, structural causal reasoning, and SHAP-based explainability, the proposed **SCDL-VFM** framework generates personalized recommendation scores for candidate fashion products. The objective of this stage is to rank all candidate items according to their relevance and recommend the **Top-K** products that best match an individual user's preferences.

Let the final feature representation obtained from the Adaptive Multimodal Fusion module be denoted by F, and let Crepresent the causal representation generated by the Structural Causal Model. These two representations are integrated through a fully connected prediction layer to compute the recommendation score.

The recommendation score is calculated as

$$\hat{y} = \sigma(W[F; C] + b)$$

where

- $\hat{y}$= Predicted recommendation probability
- W= Trainable weight matrix
- b= Bias vector
- $\sigma(\cdot)$= Sigmoid activation function

The predicted scores of all candidate products are sorted in descending order, and the **Top-K** items are recommended to the user.

$$TopK = Rank(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$$

where Ndenotes the total number of candidate products.

Training Objective

The proposed framework is trained using the **Binary Cross-Entropy (BCE)** loss function, which minimizes the difference between predicted recommendations and the actual user interactions.

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where

- y_i= Ground-truth interaction (purchase/click)
- $\hat{y}_i$= Predicted probability

The model parameters are optimized using the **Adam optimizer**, enabling efficient convergence and stable learning.

Algorithm 2: Recommendation Generation

Input

- User behaviour U
- Product image I
- Product description D
- AI-generated virtual model V_m

Output

- Top-K personalized recommendations

Steps

1. Extract visual embeddings using CLIP.
2. Encode textual descriptions using BERT.
3. Generate user behavioural embeddings.
4. Apply Adaptive Multimodal Fusion.
5. Estimate causal relationships using SCM.
6. Compute SHAP-based explanations.
7. Predict recommendation probability.
8. Rank products according to prediction scores.
9. Return the **Top-K** recommended fashion products.

Computational Complexity

The computational complexity of the proposed framework is primarily determined by the CLIP encoder, BERT encoder, Adaptive Multimodal Fusion, and Structural Causal Modeling modules. During inference, recommendation generation requires a single forward pass through the network, making the framework suitable for deployment in large-scale fashion e-commerce environments. Furthermore, the multimodal fusion and causal reasoning modules introduce only a modest computational overhead while significantly improving recommendation robustness and interpretability.

Output of the Proposed Framework

The proposed SCDL-VFM framework produces:

- Personalized Top-K fashion recommendations.

- Recommendation probability for each candidate item.
- Causal influence analysis.
- SHAP-based feature explanations.
- Explainable and trustworthy recommendation results.

4. Experimental Setup

4.1 Dataset Description

The proposed SCDL-VFM framework is evaluated using publicly available fashion datasets to ensure reproducibility and fair comparison with existing recommendation models. DeepFashion2 is employed as the primary dataset because it contains diverse clothing categories, rich attribute annotations, clothing landmarks, segmentation masks, and consumer-to-shop image pairs. To further evaluate recommendation performance in a real-world e-commerce environment, user interaction data from the H&M Personalized Fashion Recommendations Dataset are incorporated. This dataset includes customer purchase history, product metadata, and transaction records, enabling personalized recommendation experiments.

Before model training, all images are resized to a fixed resolution and normalized. Product descriptions are cleaned by removing special characters and stop words, while user interaction data are transformed into behavioral feature vectors. Missing values are handled through standard preprocessing techniques.

4.2 Experimental Environment

The proposed framework is implemented using **Python 3.11** and the **PyTorch** deep learning library. Model training is performed on **Google Colab Pro** equipped with an **NVIDIA Tesla T4 GPU**.

Component	**Specification**
Programming Language	Python 3.11
Deep Learning Framework	PyTorch
Visual Encoder	CLIP
Text Encoder	BERT
Optimizer	Adam
GPU	NVIDIA Tesla T4
Operating Environment	Google Colab Pro

4.3 Hyperparameter Settings

The hyperparameters are selected based on preliminary experiments and commonly adopted settings in multimodal recommendation research.

Parameter	Value
Batch Size	64
Learning Rate	0.0001
Epochs	50
Embedding Dimension	512
Optimizer	Adam
Dropout	0.3
Top-K	10

4.4 Evaluation Metrics

To comprehensively evaluate recommendation performance, multiple ranking-based evaluation metrics are employed.

- **Precision@K** – Measures the proportion of relevant items among the Top-K recommendations.
- **Recall@K** – Measures the percentage of relevant items successfully retrieved.
- **NDCG@K** – Evaluates ranking quality by considering the positions of relevant items.
- **MAP (Mean Average Precision)** – Measures the average precision across all users.
- **MRR (Mean Reciprocal Rank)** – Evaluates how early the first relevant recommendation appears.
- **Hit Rate (HR@K)** – Measures whether at least one relevant item appears in the Top-K recommendations.

These metrics provide a comprehensive assessment of both recommendation accuracy and ranking effectiveness.

4.5 Baseline Models

The proposed **SCDL-VFM** framework is compared with the following state-of-the-art recommendation models:

Model	Description
NCF	Neural Collaborative Filtering
DeepFM	Deep Factorization Machine
BERT4Rec	Transformer-based Sequential Recommendation
CLIP-based Multimodal Model	Vision-language recommendation
Causal Recommendation Model	SCM-based recommendation
SCDL-VFM (Proposed)	Multimodal + SCM + SHAP

6. CONCLUSION

This paper presented **SCDL-VFM**, a novel **Structural Causal Explainable Deep Learning Framework** for personalized fashion recommendation using AI-generated virtual fashion models. The proposed framework integrates multimodal feature extraction, Adaptive Cross-Modal Fusion (ACMF), the Causal Preference Refinement Network (CPRN), Structural Causal Modeling (SCM), and SHAP-based explainability into a unified recommendation architecture.

Unlike conventional recommendation systems that primarily rely on correlation-based learning, the proposed framework incorporates causal reasoning to better capture the underlying relationships between user preferences, product attributes, and recommendation outcomes. The integration of multimodal information, including visual features, textual descriptions, user behavioral patterns, and AI-generated virtual fashion models, is expected to improve recommendation relevance while enhancing model robustness and personalization.

Furthermore, the incorporation of SHAP-based explanations provides interpretable recommendation decisions, thereby increasing transparency and user trust. The modular architecture of SCDL-VFM also enables flexible integration with future recommendation technologies and large-scale fashion e-commerce platforms.

Overall, the proposed framework establishes a comprehensive theoretical foundation for next-generation explainable and causally informed fashion recommendation systems. Future experimental validation on publicly available datasets will further verify its effectiveness using standard recommendation evaluation metrics.

Future Work

Several promising research directions can further enhance the proposed SCDL-VFM framework.

Future work will focus on implementing the complete framework using publicly available fashion recommendation datasets such as **DeepFashion2** and the **H&M Personalized Fashion Recommendations Dataset** to perform comprehensive experimental validation. Performance will be evaluated using standard recommendation metrics, including Precision@K, Recall@K, NDCG@K, MAP, MRR, and Hit Rate.

The framework can also be extended by incorporating transformer-based multimodal architectures, graph neural networks for user-item relationship modeling, reinforcement learning for adaptive recommendation, and large language models for conversational fashion assistance. Additionally, future research may explore federated learning and privacy-

preserving recommendation techniques to protect user data while maintaining high recommendation quality.

Another promising direction involves integrating real-time user feedback and online learning mechanisms to continuously refine customer preferences. The proposed causal reasoning framework may also be adapted for other personalized recommendation domains such as cosmetics, furniture, healthcare products, and digital retail platforms.

Overall, these future enhancements have the potential to improve the scalability, interpretability, robustness, and practical applicability of intelligent recommendation systems in modern e-commerce environments.

REFERENCES

1. F. Ricci, L. Rokach, and B. Shapira, *Recommender Systems Handbook*, 3rd ed. Cham, Switzerland: Springer, 2022.
2. X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. S. Chua, "Neural Collaborative Filtering," in *Proceedings of the 26th International Conference on World Wide Web (WWW)*, Perth, Australia, 2017, pp. 173–182.
3. Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems," *IEEE Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.
4. Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 1096–1104.
5. Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-identification of Clothing Images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 5337–5345.
6. X. Han, Z. Wu, Y. Jiang, and L. S. Davis, "Learning Fashion Compatibility with Bidirectional LSTMs," in *Proceedings of the ACM International Conference on Multimedia*, Mountain View, CA, USA, 2017, pp. 1078–1086.
7. M. Vasileva, B. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth, "Learning Type-Aware Embeddings for Fashion Compatibility," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 390–405.
8. A. Vaswani *et al*., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

9. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
10. A. Radford *et al*., "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, Virtual Event, 2021.
11. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 1135–1144.
12. S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4765–4774.
13. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
14. J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2009.
15. B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward Causal Representation Learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.
16. L. Bottou, J. Peters, J. Quiñonero-Candela, D. Charles, D. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3207–3260, 2013.
17. W. Wang, F. Feng, X. He, H. Wang, and T. S. Chua, "Causal Recommendation: Progresses and Future Directions," *ACM Transactions on Information Systems (TOIS)*, vol. 40, no. 4, pp. 1–38, 2022.
18. C. Gao, X. Wang, X. He, and Y. Li, "Causal Inference in Recommender Systems: A Survey and Future Directions," *ACM Transactions on Information Systems (TOIS)*, vol. 41, no. 3, pp. 1–38, 2023.