

Article

HAIRec: A Hybrid Recommendation Framework Integrating Review Sentiment and Interaction Information

Ganglong Duan and Tianqiao Gong * 

Department of Economics and Management, Xi'an University of Technology, Xi'an 710048, China;
gl-duan@xaut.edu.cn

* Correspondence: 2240520069@stu.xaut.edu.cn; Tel.: +86-153-19454279

Abstract

Recommender systems are widely used in e-commerce to model user preferences and deliver personalized services. However, conventional interaction-based recommendation approaches fail to fully capture users' underlying motivations and emotional tendencies. User-generated reviews offer complementary sentiment information, yet integrating sentiment semantics with interaction modeling remains challenging due to feature heterogeneity and sparsity. This paper proposes a hybrid recommendation framework (HAIRec) that jointly models review sentiment and interaction information within a unified architecture. The model extracts sentiment representations, latent rating factors, user preference features, and item attributes through multi-source feature learning. To address the semantic heterogeneity between review sentiment and interaction behavior, a hierarchical heterogeneous feature fusion architecture is proposed. Specifically, a dual-level attention mechanism is designed to jointly model global feature importance and fine-grained semantic interactions, designed to facilitate multi-granularity preference learning across heterogeneous domains. In addition, a Deep & Cross Network V2 (DCN-V2) is adopted to explicitly and implicitly model high-order feature interactions. Extensive experiments are conducted on five real-world subcategories of the Amazon Product Dataset (including Movies and TV, Digital Music, Musical Instruments, Toys and Games, and Home and Kitchen). The experimental results demonstrate that the proposed HAIRec framework consistently outperforms several representative baselines (such as DeepCoNN, NARRE, and LightGCN). Notably, compared with the strong LightGCN benchmark, HAIRec achieves significant rating prediction improvements, reducing the Mean Squared Error (MSE) to 0.884 on Movies and TV, 0.882 on Digital Music, and 0.863 on Toys and Games, which corresponds to maximum performance gains of up to 7.14% in MSE reduction.

Keywords: recommender systems; sentiment analysis; multi-level attention; user preference



Academic Editor: George
Angelos Papadopoulos

Received: 18 June 2026

Revised: 4 August 2026

Accepted: 4 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Driven by the rapid development of digital infrastructure and mobile connectivity, online commerce has become a primary channel for consumers to access product information and complete transactions, generating massive amounts of user behavior data and personalized service demands [1]. In this context, recommender systems have evolved into a core component of modern e-commerce platforms, serving as an effective solution for alleviating information overload and delivering personalized services [2]. By modeling user-item interactions, recommender systems reduce users' decision-making costs while improving recommendation quality and platform efficiency [3].

Despite their remarkable success, traditional recommendation approaches, such as collaborative filtering and content-based methods, still face several challenges in complex real-world scenarios. First, data sparsity remains a critical issue because most users interact with only a small fraction of available items, resulting in unreliable preference estimation, especially for long-tail products [4]. Second, the cold-start problem significantly limits recommendation accuracy for newly introduced users and items due to the absence of historical interaction records [5]. Third, traditional models mainly rely on explicit feedback and therefore struggle to capture users' implicit preferences, such as emotional tendencies and nuanced demand intensity, leading to limited recommendation interpretability [6].

To address these limitations, recommendation research has increasingly incorporated user-generated reviews as an additional source of heterogeneous information. As an important task in natural language processing, sentiment analysis enables the extraction of emotional signals [7] and latent user preferences from review texts [8], thereby providing complementary semantic information beyond user–item interactions. Consequently, review-aware recommendation methods have emerged to jointly exploit interaction records and review texts for more comprehensive user preference modeling. However, review-based models remain constrained by review sparsity and semantic ambiguity, whereas interaction data mainly capture users' macro-level behavioral preferences but lack the ability to explain the underlying motivations behind user decisions [9]. Consequently, relying solely on either textual reviews or interaction information is insufficient for comprehensively modeling user preferences.

Despite the progress achieved by existing review-aware recommendation methods, they still face a fundamental challenge: the semantic granularity mismatch and insufficient high-order feature interaction modeling between heterogeneous information sources under sparse data conditions. Specifically, interaction features mainly reflect coarse-grained behavioral preferences, whereas review texts contain fine-grained emotional and contextual semantics. Most existing fusion strategies rely on simple feature concatenation or shallow attention mechanisms, making them less effective for accurate cross-granularity semantic alignment, while posing challenges in capturing complex high-order feature interactions between heterogeneous sources.

To better model user behaviors and contextual dependencies, transformer-based recommendation models, such as SASRec [10] and BERT4Rec [11], have been widely adopted. Although these models demonstrate strong sequential representation learning capability, they primarily operate within a unified semantic space, which may limit their capability to fully capture heterogeneous semantic complementarity under sparse recommendation scenarios. Meanwhile, graph-based recommendation models leverage user–item interaction graphs to learn collaborative relationships through information propagation. Although effective for structural representation learning, existing graph propagation mechanisms are susceptible to noisy information propagation and over-smoothing under sparse interaction conditions [12], while still insufficiently modeling the fine-grained emotional semantics embedded in review texts [13].

To help address these limitations, this work proposes a Hierarchical Heterogeneous Interaction-Aware Recommendation framework (HAIRec) for sparse review-aware recommendation tasks. Unlike existing recommendation models that mainly emphasize either sequential semantic modeling or structural collaborative propagation independently, HAIRec jointly models fine-grained emotional semantics and coarse-grained behavioral preferences through hierarchical heterogeneous feature interaction learning [14].

Specifically, a hierarchical feature fusion architecture is developed to progressively integrate heterogeneous representations, while a dual-level attention mechanism adaptively captures both global module importance and fine-grained semantic dependencies. Fur-

thermore, explicit and implicit high-order feature interaction learning are jointly employed to strengthen cross-domain semantic alignment and heterogeneous representation learning, thereby improving recommendation accuracy, robustness, and interpretability under sparse recommendation scenarios. The main contributions of this work are summarized as follows:

1. A hierarchical hybrid recommendation architecture is proposed to jointly model heterogeneous review sentiment and user–item interaction information, aiming to alleviate the semantic granularity inconsistency between textual and behavioral features.
2. A dual-level attention mechanism is designed to simultaneously capture global module importance and fine-grained semantic dependencies, enabling adaptive cross-domain feature fusion and improving heterogeneous preference representation learning.
3. A heterogeneous feature interaction framework combining explicit and implicit high-order interaction learning is introduced to model complex cross-domain semantic dependencies, thereby enhancing semantic interaction learning, recommendation robustness, and prediction accuracy under sparse recommendation scenarios.

2. Related Work

2.1. Sentiment Analysis

Sentiment analysis, an essential subfield of natural language processing (NLP), focuses on computationally detecting, extracting, and classifying subjective emotional traits in textual data. It has been widely applied to review data, helping identify emotional tendencies such as positive, neutral, or negative sentiments [15]. Early paradigms in this domain heavily relied on lexicon-based methods, which utilize predefined emotional dictionaries (such as SentiWordNet or WordNet) and syntactic rules to aggregate text valence [16]. While highly interpretable and requiring no extensive training data, these methods typically struggle with contextual variations and out-of-vocabulary expressions.

To capture more localized and granular opinion targets, aspect-based sentiment analysis (ABSA) emerged as a crucial finer-grained paradigm. Rather than assigning a single sentiment score to an entire review, ABSA decomposes textual feedback into specific product features (aspects) and identifies the corresponding target-specific sentiments, providing a nuanced understanding of multifaceted user evaluations [17].

Recently, the sentiment analysis landscape has fundamentally shifted toward approaches built on pretrained language models (PLMs), such as BERT, RoBERTa, and their variants. By exploiting deep self-attention mechanisms and bidirectional contextual representation learning, PLMs successfully resolve complex long-range semantic dependencies and implicit linguistic nuances, including intense irony, sarcasm, or complex negations. In modern recommender systems, leveraging PLMs to extract fine-grained semantic profiles from user reviews has established a robust foundation for capturing latent preference dynamics and bridging the semantic gap between behavioral logs and textual feedback [18].

2.2. Review-Aware Recommendation Models

Recommender systems aim to predict user preferences by modeling historical user–item interactions, such as ratings, clicks, and purchase behaviors. Collaborative filtering and content-based methods are widely adopted, while hybrid recommendation frameworks further improve personalization by integrating multiple information sources.

To alleviate sparsity and cold-start issues, recent review-aware recommendation models increasingly incorporate textual semantic information from user reviews. As summarized in Table 1, representative methods such as PMF [19], HFT [20], DeepCoNN [21], ConvMF [22], D-Attn [23], and NARRE [24] improve recommendation performance through

collaborative filtering, topic modeling, convolutional semantic extraction, or attention-based review representation learning.

Table 1. Comparison of representative review-aware recommendation models.

Model	Review Modeling	Feature Fusion Strategy	Attention Modeling	Heterogeneous Semantic Interaction
PMF	None	None	None	None
HFT	Topic Modeling	Latent Factor Alignment	None	Weak
DeepCoNN	CNN	Feature Concatenation	None	Limited
ConvMF	CNN + MF	Shallow Fusion	None	Limited
D-Attn	Attention-based Review Representation	Attention Fusion	Single-level	Partial
NARRE	Review Attention	Attention Fusion	Single-level	Partial
ANR	Aspect-based CNN	Matrix Factorization Alignment	Multi-level	Partial
MPCN	Hierarchical Pointer	Gating Mechanism	Multi-level	Partial
CARL	Contextual CNN	Layer-wise Projection	Single-level	Partial
DAML	Local-Global Joint CNN	Neural Factorization Machine	Multi-level	Partial
HAIRec	Multi-granularity Semantic Modeling	Hierarchical Heterogeneous Fusion	Dual-level	Full

Specifically, PMF [19] learns latent user and item representations through probabilistic matrix factorization and has become a classical benchmark for collaborative filtering. However, it relies solely on user–item interaction data and cannot exploit textual semantic information, making it vulnerable to sparse and cold-start scenarios. HFT [20] successfully regularizes latent profiles by aligning topic distributions with rating dimensions. However, it relies on a bag-of-words assumption that largely overlooks the structural and word-order dependencies within reviews. To capture contextual sequences, DeepCoNN [21] effectively extracts non-linear representations using dual-channel convolutional neural networks, and ConvMF [22] integrates convolutional layers into probabilistic matrix factorization to better handle sparse item metadata. Although these deep models successfully capture local word-level features, their primary limitation is that they treat all text tokens with equal importance, which propagates significant noise from uninformative reviews. D-Attn [23] further introduces a dual-attention mechanism to better identify informative semantic representations from review texts. Nevertheless, its attention mechanism mainly focuses on textual information and lacks explicit modeling of heterogeneous interactions between review semantics and user–item behavioral features. To further improve review utilization, NARRE [24] introduces a review-level attention mechanism to adaptively estimate the usefulness of individual reviews.

Beyond these foundational deep models, recent advanced studies explore more sophisticated multi-level or dual attention mechanisms to capture review semantics at finer granularities. For instance, ANR [25] introduces an aspect-based neural recommendation framework that designs an aspect-level attention component to model user-item preferences across distinct latent aspects. MPCN [26] leverages a hierarchical multi-pointer co-attention network to dynamically select critical words and reviews from user-item review banks. CARL [27] utilizes a context-aware review learning approach that applies dynamic local attention to map contextual review features into a low-dimensional space.

Similarly, DAML [28] adopts a dual attention mutual learning paradigm, jointly optimizing local and mutual attention components to align rating interactions with text semantics.

These representative methods were selected as baseline models because they cover the major development stages of review-aware recommendation, ranging from collaborative filtering (PMF) and topic modeling (HFT) to convolution-based semantic learning (DeepCoNN and ConvMF) and attention-based review modeling (D-Attn and NARRE). Their diverse modeling strategies provide comprehensive and representative benchmarks for evaluating the effectiveness of the proposed HAIRec framework.

Nevertheless, as summarized in Table 1, existing hybrid recommendation approaches (e.g., DeepCoNN, ANR, and DAML) still face several limitations. Most of these methods primarily rely on single-level attention mechanisms or relatively shallow semantic fusion strategies, such as feature concatenation and conventional interaction modeling, which are insufficient to capture fine-grained semantic dependencies among heterogeneous features and complex high-order non-linear interactions across different representation spaces.

To address these limitations and further clarify the methodological position of this work, HAIRec is designed with two complementary components: a hierarchical dual-level attention network that adaptively captures local sentiment-level semantics and global feature importance, and an integrated Deep & Cross Network (DCN)-based interaction module that explicitly models high-order feature interactions between heterogeneous representations. By jointly incorporating these components, HAIRec enables dynamic alignment and interaction between multi-granularity review semantics and user–item behavioral information, providing an effective solution for improving preference representation under sparse recommendation scenarios.

2.3. Graph-Based Recommendation Models

Graph-based recommendation models have attracted extensive attention due to their strong capability in modeling high-order collaborative relationships through graph propagation mechanisms. Representative methods such as NGCF [29] and LightGCN [30] construct user–item interaction graphs to capture collaborative preference dependencies and structural connectivity patterns. By propagating neighborhood information over graph structures, these models effectively enhance collaborative filtering performance under sparse interaction conditions. However, existing graph-based recommendation methods mainly focus on structural interaction learning and often lack fine-grained emotional semantic modeling from review texts. In sparse recommendation scenarios, graph propagation may also introduce noisy neighbor information, which can weaken semantic discriminability and recommendation robustness.

2.4. LLM-Based and Retrieval-Augmented Recommendation

Recently, Large Language Model (LLM)-based recommendation methods and retrieval-augmented recommendation frameworks have demonstrated strong semantic reasoning and contextual understanding capabilities. By leveraging large-scale pretrained language models, these approaches can capture complex user intent and semantic relationships from textual content. Typically, LLM-based approaches adapt generative paradigms to recommendation tasks either by fine-tuning open-source LLMs with behavioral data, as demonstrated in TallRec [31], or by utilizing prompt engineering with closed-source models to formulate recommendations as natural language generation, such as Chat-Rec [32]. These methods excel at zero-shot generalization and utilizing open-world knowledge to interpret vague user preferences.

Simultaneously, retrieval-augmented generation (RAG) recommendation frameworks further improve recommendation relevance through external memory retrieval and se-

mantic matching mechanisms. For instance, frameworks like A-RAG [33] decouple the recommendation process into a dense retrieval phase—which surfaces relevant historical reviews or candidate items from an external corpus—and a generation phase that synthesizes these contexts. This helps reduce the inherent hallucinations of LLMs and offers more traceable rationales for recommendations.

Nevertheless, existing LLM-based recommendation architectures generally require substantial computational resources and large-scale pretraining data, suffering from high inference latency during real-time serving. In addition, most current approaches primarily emphasize semantic reasoning or retrieval enhancement independently. They often project textual features and behavioral interactions into separate spaces or rely on simple serialization, meaning that the multi-layered heterogeneous interaction learning between fine-grained review sentiment and explicit behavioral preference information remains insufficiently explored.

3. Model Design

3.1. Hybrid Recommendation Framework Based on Sentiment and Interaction Information

Existing recommendation frameworks often suffer from insufficient semantic alignment between user review representations and interaction-based preference modeling. Textual reviews contain fine-grained emotional semantics, while interaction data mainly reflect coarse-grained behavioral patterns. Direct naive aggregation of these heterogeneous features may introduce semantic inconsistency and information redundancy, limiting the effectiveness of feature interaction learning. To address this issue, the proposed HAIRec framework adopts a hierarchical heterogeneous feature fusion strategy, jointly modeling multi-source review sentiment and interaction information within a unified architecture as illustrated in Figure 1.

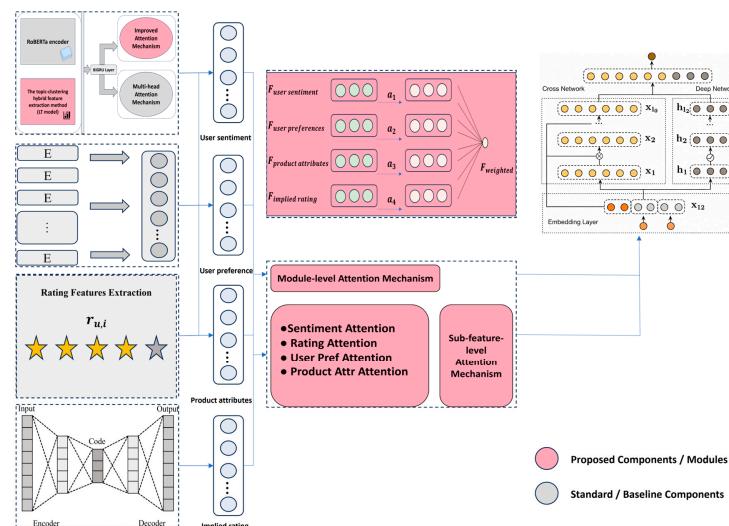


Figure 1. Operational pipeline of the proposed HAIRec framework. Solid arrows denote directed tensor forward propagation. **(Left):** Parallel extraction of heterogeneous features, projected into dense hidden vectors (blue circles). **(Center):** Hierarchical feature fusion dynamically weighting representations via dual-level attention (comprising sub-feature-level and module-level attention) to filter noise. **(Right):** High-order interaction modeling via the DCN-V2 module, consisting of parallel Cross Network and Deep Network components, to capture high-order feature interactions and output the predicted rating. (The visual illustration of the DCN-V2 sub-diagram on the right is adapted from Wang et al. [34]).

The overall hierarchical information flow of HAIRec is coordinated through tightly integrated processing stages within a cohesive system. Initially, heterogeneous representa-

tions are simultaneously extracted from four complementary semantic streams. Raw review texts are processed by the RoBERTa–BiGRU sentiment module to extract sentiment-aware semantics. Crucially, the BiGRU architecture is deliberately selected over standard LSTM or regular GRU because it simultaneously captures bidirectional contextual dependencies from review texts without losing long-range semantic modifiers [35], while omitting the dedicated output gate found in LSTM to maintain a streamlined parameter profile that minimizes computational latency and mitigates overfitting risks under data-sparse settings [36]. Concurrently, sparse rating matrices are projected into dense low-dimensional features via a deep autoencoder, while user preference profiles and item attribute representations are constructed in parallel utilizing historical behaviors and metadata descriptions.

Following this multi-source feature extraction, the derived heterogeneous representations are seamlessly integrated through a proposed hierarchical dual-level attention mechanism. This module adaptively operates at both the sub-feature level (refining local semantic dependencies) and the module level (weighting global inter-domain feature importance), effectively filtering out uninformative textual noise to amplify highly correlated signal correlations. Subsequently, the fused joint representation is fed into DCN-V2 [34] to capture both explicit non-linear cross-features and implicit high-order correlations in parallel, promoting a more comprehensive and expressive feature interaction learning process. Finally, the resulting high-order feature representation is processed by an optimized prediction layer to output the estimated rating score for the target user–item pair.

3.2. Motivation for Hierarchical Attention Design

The proposed framework does not aim to redesign existing backbone neural components independently. Instead, the core contribution lies in the hierarchical organization and collaborative interaction modeling of heterogeneous semantic representations extracted from multiple complementary feature domains. In hybrid recommendation scenarios, heterogeneous information originates from different semantic granularities and functional domains. Specifically, interaction features mainly reflect coarse-grained behavioral preferences, whereas review representations contain fine-grained emotional and contextual semantics. Traditional single-level attention mechanisms typically perform feature weighting within a unified semantic space, which may lead to semantic information dilution and insufficient cross-domain dependency modeling when heterogeneous features exhibit substantial semantic differences.

To address this limitation, the proposed framework introduces a hierarchical dual-level attention mechanism for feature interaction and fusion. The module-level attention dynamically models the global importance of sentiment, rating, preference, and item representations through adaptive Softmax weighting, while the sub-feature attention further captures fine-grained semantic dependencies within heterogeneous feature spaces. This hierarchical decomposition separates local semantic extraction from global feature aggregation, thereby reducing semantic interference among heterogeneous domains and improving discriminative representation learning.

For a conventional single-level attention mechanism, heterogeneous feature representations are directly aggregated within a shared semantic space:

$$u = \sum_{i=1}^n \alpha_i h_i \quad (1)$$

where h_i denotes the feature representation and α_i represents the normalized attention weight. This formulation performs direct global feature aggregation by assigning different importance weights to heterogeneous representations within a unified semantic space. However, when heterogeneous features exhibit substantial semantic differences, such

direct aggregation may suppress local semantic structures, weaken cross-domain semantic discriminability, and limit fine-grained preference interaction learning.

In the proposed framework, feature interaction is decomposed into hierarchical stages. First, sub-feature attention computes fine-grained semantic representations within each feature module:

$$r_j = \sum_{i=1}^{n_j} \alpha_{ij} h_{ij} \quad (2)$$

where h_{ij} denotes the i -th sub-feature of module j , and r_j represents the refined module representation. Subsequently, module-level attention performs global feature aggregation across heterogeneous domains:

$$u = \sum_{j=1}^m \beta_j r_j \quad (3)$$

where β_j denotes the module-level attention weight.

Compared with conventional single-level attention strategies, the proposed hierarchical attention mechanism separates local semantic extraction from global feature interaction. This hierarchical decomposition reduces semantic interference among heterogeneous domains and improves the preservation of discriminative contextual information.

Furthermore, a DCN is employed to jointly model explicit feature crossing and implicit nonlinear semantic interactions, enabling more effective high-order heterogeneous feature learning. Finally, the prediction score is generated through a multilayer perceptron (MLP).

3.3. Feature Extraction Module

The proposed framework performs hierarchical feature learning through four stages.

First, the Latent Topic (LT) module and RoBERTa encoder jointly extract topic-level semantic distributions and contextual semantic representations from review texts.

Second, the BiGRU and improved attention mechanism further model sequential dependencies and sentiment salience to generate fine-grained emotional representations.

Third, the autoencoder extracts latent collaborative preference features from sparse user–item interaction matrices, while statistical preference features and item attribute representations provide complementary behavioral and structural information.

Finally, the extracted representations are integrated through the proposed hierarchical dual-level attention mechanism and DCN-based interaction learning framework to jointly model global feature importance and high-order semantic dependencies for rating prediction.

3.3.1. Emotional Feature Extraction from Comments

To quantify textual sentiment tendencies and extract multi-level semantic features, a Latent topic-RoBERTa-BiGRU-Preference (LRBP)-based model is developed. The model takes historic review collections (excluding the target review) as input, employing an LT-based topic modeling module and a RoBERTa encoder to extract complementary thematic and contextual semantic representations, followed by a bidirectional GRU (BiGRU) to capture sequential dependencies [37]. Based on this, a dual-channel architecture is designed: one channel applies an improved attention mechanism with a Sigmoid activation to generate sentiment scores, while the other utilizes a multi-head attention mechanism to capture user preferences and attribute-related features for sentiment-aware preference modeling [38].

- Aspect-based Semantic Encoder (LT Model)

To extract global thematic preference structures across the review corpus, a topic-clustering hybrid feature extraction method (LT model) is introduced. The model employs Latent Dirichlet Allocation (LDA) to learn corpus-level topic distributions through a probabilistic document–topic–word framework optimized via Gibbs sampling [39]. Unlike

RoBERTa, which captures sentence-level contextual semantics, the LDA-based topic modeling focuses on global thematic granularity, providing complementary semantic information. To reduce redundancy and noise before feature fusion, K-means clustering and Principal Component Analysis (PCA) are applied to project the converged high-dimensional topic vectors into a unified 768-dimensional semantic space.

- **RoBERTa Layer**

To efficiently capture deep semantic information, a pre-trained RoBERTa model is employed to encode review texts. RoBERTa maps the input text to a series of context-dependent word vector representations. Instead of being used directly for prediction, these embeddings are element-wise combined with the LT topic features to enhance the expressive power of sentiment cues and avoid over-reliance on isolated emotional words. The input sequence is transformed into a context-aware representation matrix $T \in \mathbb{R}^{n \times 768}$ through standard Transformer encoder layers.

- **BiGRU Layer**

A BiGRU is deployed to model contextual sequential features from the fused topic-semantic embeddings. By simultaneously encoding sequences from both forward and backward directions, the BiGRU captures comprehensive contextual semantic information through concatenated hidden states, ensuring robust sequence modeling under data-sparse recommendation scenarios.

- **Improved Attention Mechanism Layer**

Standard attention mechanisms often suffer from “attention bias” or cognitive tunneling, where hyper-emotional outlier tokens dominate the Softmax gradient paths. This optimization bottleneck causes the model to over-concentrate weights on isolated words while neglecting critical contextual indicators, ultimately weakening the semantic diversity and performance of recommendation modeling [40].

As schematically illustrated in Figure 2, the standard attention allocation spikes abnormally on isolated emotional words, heavily diluting the surrounding contextual semantics. In contrast, Figure 3 illustrates that the proposed improved attention mechanism effectively redistributes attention weights across both salient emotional tokens and contextual semantic cues, thereby preserving richer semantic information.

To alleviate this bottleneck, the proposed improved mechanism introduces a dynamic masking strategy acting as a self-supervised regularization constraint. Two dynamic sample-specific token sets are formally maintained: a positive semantic set Q_p containing tokens that exert active and constructive impacts on preference prediction, and a misleading semantic set Q_n that captures tokens triggering severe cognitive tunneling and optimization bias. The formal pipeline proceeds through four sequential steps:

- (1) **Token Representation Generation:** Given the dynamic parameters θ , the global category anchor representation is generated via $A = f_a(\theta)$. To eliminate historical attention noise, the current textual feature matrix X undergoes proactive masking configurations filtered by Q_p and Q_n , yielding the refined text matrix X' is formulated as:

$$X' = X \setminus (Q_p \cup Q_n) \quad (4)$$

Standard Attention Weight Distribution (Schematic)

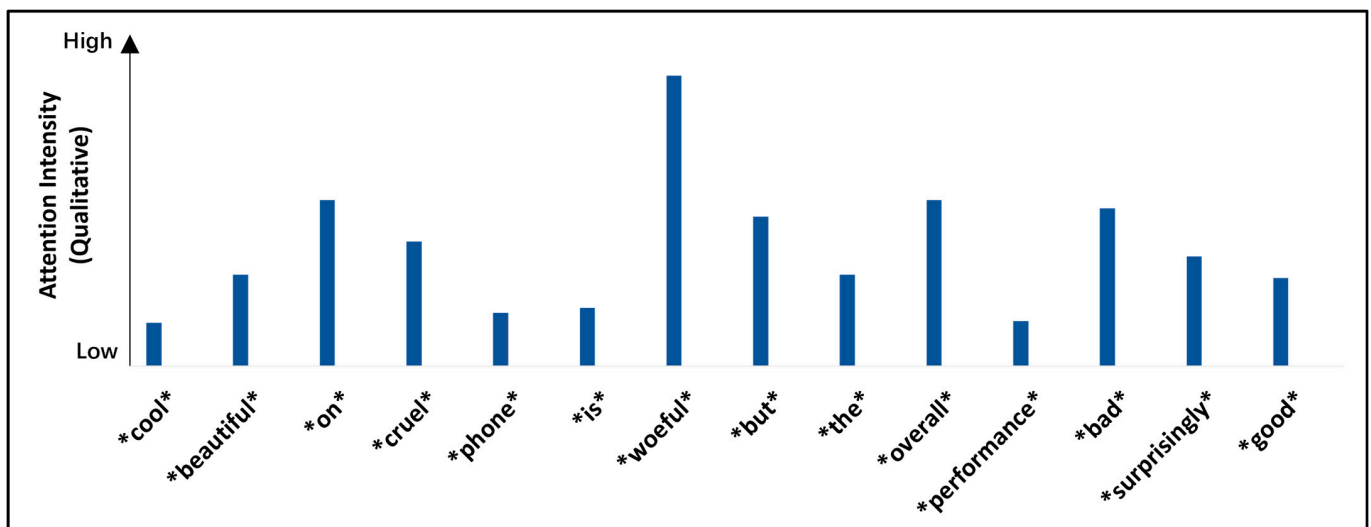


Figure 2. Schematic diagram illustrating the attention weight distribution under the standard attention mechanism (qualitative representation, not actual model output).

Improved Attention Weight Distribution (Schematic)

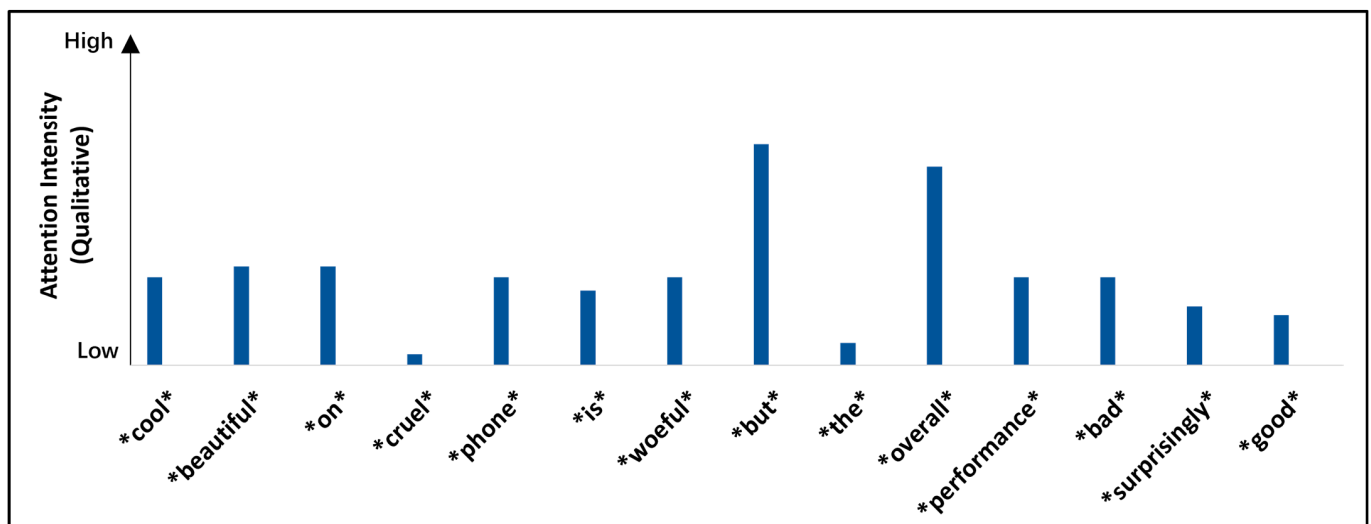


Figure 3. Schematic diagram illustrating the attention weight distribution under the proposed improved attention mechanism (qualitative representation, not actual model output).

The corresponding hidden contextual token representation H is sequentially mapped via a parameter network:

$$H = f_{word}(X') \quad (5)$$

- (2) Preference Classification and Attention Weighting: The prediction routing network estimates the final score mapping $\hat{y}$ from the hidden code space H via Equation (6):

$$\hat{y} = f_{prediction}(H) \quad (6)$$

Concurrently, the relative semantic contribution matrix α for each individual token is extracted through the parameter layer and is formulated as:

$$\alpha = f_{attention}(H) \quad (7)$$

- (3) **Dynamic Thresholding and Token Set Upgrading:** To adapt to varying sentence granularities, the global expectation baseline of all token attention weights is calculated as $E(\alpha)$. A dynamic decision threshold is parameterized based on sequence length and dictionary frequency and is formulated as:

$$\text{threshold} = f(\text{length}, \text{frequency}) \quad (8)$$

A joint gating mechanism systematically monitors whether high-weight outlier tokens skew the distribution profiles. If the batch behavior triggers an anomaly ($E(\alpha) \geq \text{threshold}$), the dominant token tracking index h_{max} is dynamically harvested via an optimization gate, which is formulated as:

$$\text{if } E(\alpha) < \text{threshold, continue; else extract } h_{max} = \underset{h \in H}{\operatorname{argmax}} \alpha(h) \quad (9)$$

Depending on whether the structural prediction $\hat{y}$ matches the ground-truth label y , the dynamic tracking sets Q_p and Q_n are adaptively updated. If the presence of h_{max} induces a substantial classification error under a tight loss constraint, its positional index is appended to Q_n to suppress its attention footprint in subsequent iterations; otherwise, it is allocated to Q_p to reinforce its semantic routing.

- (4) **Parameter Convergence via Multi-Tuple Feedbacks:** The regularized inputs, tracking tokens, and prediction outcomes are packed as a unified self-supervised interaction tuple (X', t, y) . This tuple is synthesized into a joint training database $D(k)$ along with the historical ground-truth profile, and is formulated as:

$$D(k) = \{(X', t, y), (\text{correct_result})\} \quad (10)$$

The synthesized corpus $D(k)$ is subsequently injected into the next backpropagation epoch to update the model weights. This architecture forces the network to explore alternative semantic propagation paths, helping mitigate localized gradient saturation, easing optimization bottlenecks, and supporting a more resilient preference representation.

3.3.2. Implicit Rating Feature Extraction

To capture global user–item relational profiles and tackle severe interaction sparsity, a deep Autoencoder (AE) framework is employed to extract latent collaborative features directly from the user–item interaction matrix $R \in \mathbb{R}^{n \times m}$, where n and m denote the numbers of users and items, respectively. The interaction matrix R is constructed exclusively from the training subset, while all test interactions are strictly excluded during autoencoder training and latent representation extraction to prevent potential data leakage. Rather than relying on simple low-rank factorizations, the non-linear autoencoding bottleneck maps sparse transactional signals into a dense, low-dimensional space to reconstruct unseen preference scores, which has been proven highly effective in advanced collaborative filtering paradigms [41,42].

The extraction mechanism follows a standard bottleneck routing strategy. The encoder channel compresses the high-dimensional sparse rating vectors through multiple nonlinear transformation layers. Each encoder layer is formulated as:

$$H_{enc}^{(l)} = \text{Dropout}\left(\sigma\left(W_{enc}^{(l)} H_{enc}^{(l-1)} + b_{enc}^{(l)}\right)\right) \quad (11)$$

where $H_{enc}^{(0)}$ denotes the input rating vector, $W_{enc}^{(l)}$ and $b_{enc}^{(l)}$ are the trainable parameters of the l -th encoder layer, $\sigma(\cdot)$ represents the nonlinear activation function (ReLU), and Dropout is adopted to reduce overfitting.

To mitigate overfitting risks under highly incomplete rating topologies, Dropout regularization is subsequently applied to these hidden states. The resulting dense column vector explicitly encodes the target user's implicit preference traits and latent behavioral habits. The decoder reconstructs the original interaction vectors through multiple nonlinear decoding layers defined as:

$$H_{dec}^{(k)} = \sigma(W_{dec}^{(k)} H_{dec}^{(k-1)} + b_{dec}^{(k)}) \quad (12)$$

where $H_{dec}^{(0)}$ denotes the latent representation generated by the encoder, and $W_{dec}^{(k)}$ and $b_{dec}^{(k)}$ are the trainable parameters of the k -th decoder layer.

Under this framework, the item-specific latent representation vector is extracted from the corresponding row of the decoder weight matrix, effectively retaining global user-item affinity topologies within the shared semantic space. Dropout regularization and Adam optimization are iteratively applied until the empirical reconstruction likelihood function converges, providing robust initialization parameters for subsequent heterogeneous cross-layers.

3.3.3. User Preference Feature Extraction

To build multidimensional user profiles for personalized recommendations, user preference features are structured across behavioral and semantic dimensions [43]. For user rating preferences, long-term satisfaction and evaluation consistency are captured by extracting the User Average Rating (UAR), Rating Standard Deviation (RSD), and High-Rating Ratio (HRR). For User Price Preferences, the Median Purchase Price (MPP) and Price Sensitivity (PS) are calculated to reflect typical consumption levels and spending variability robustly. Concurrently, for User Purchase Category Preferences, a pre-trained RoBERTa model encodes normalized product category paths, outputting a global deep semantic representation C_u after L_2 normalization to enrich directional semantic affinity.

To eliminate scale inconsistency among heterogeneous domains (e.g., rating range [1, 5] versus price range [1, 10^5]) and help stabilize feature interaction, numerical features are normalized to the [0, 1] interval via min-max scaling, while semantic vectors are standardized through L_2 normalization to constrain feature magnitude variations [44]. To further align heterogeneous feature spaces, the category semantic vectors are first compressed using PCA, retaining 95% of the original variance, and subsequently projected through a fully connected layer to obtain a fixed five-dimensional semantic representation.

Formally, let

$$F_{num} = [UAR_u, RSD_u, HRR_u, MPP_u, PS_u] \in \mathbb{R}^5 \quad (13)$$

denote the normalized numerical behavioral feature vector, and let

$$V_{embed} \in \mathbb{R}^5 \quad (14)$$

represent the processed category semantic embedding after PCA and fully connected projection. Rather than directly concatenating the two feature groups with equal importance, an adaptive gating mechanism is employed to learn their relative contributions before feature fusion. The final composite user preference representation is formulated as

$$p_u = [\alpha \cdot F_{num} \parallel (1 - \alpha) \cdot V_{embed}] \quad (15)$$

where $\parallel$ denotes the tensor combination operator, and $\alpha \in [0, 1]$ is a learnable scalar gating coefficient that is dynamically estimated by

$$\alpha = \text{Sigmoid}(W_a \cdot [F_{num} \parallel V_{embed}] + b_a) \quad (16)$$

where $W_a \in \mathbb{R}^{1 \times 10}$ and $b_a \in \mathbb{R}$ are the learnable weight matrix and bias term of the gating layer, respectively.

Through end-to-end optimization, the adaptive gate automatically adjusts the relative contributions of numerical behavioral statistics and semantic category representations according to the training objective. This adaptive fusion strategy enables the model to learn a more robust and discriminative user preference representation, thereby improving feature interaction quality and enhancing recommendation performance under sparse interaction scenarios.

3.3.4. Product Attribute Feature Extraction

- Product Metadata Attributes Extraction

Product metadata features project crucial discrete item information into low-dimensional dense factor spaces to capture latent similarities [45]. Product IDs and brand names are processed through embedding layers to automatically learn and update dense vector representations during training. Product price is transformed via a ranking-based method to compute its position quantile within the global price distribution, while category paths are tokenized and encoded via RoBERTa to yield the final category feature vector E_{cat} . Following feature space alignment and normalization, all constructed embedding vectors are integrated to obtain the final comprehensive metadata feature vector X_{meta} through concatenation.

- Product Description Feature Extraction

To extract multi-dimensional attribute features from textual descriptions, this study constructs an enhanced text analysis model integrating a CNN with a Self-Attention mechanism, as illustrated in Figure 4.

The processed description text is transformed into word embeddings via RoBERTa to form the base matrix $E \in \mathbb{R}^{n \times 768}$. A 3-layer convolutional structure (Conv-1 to Conv-3) with a kernel width of 3 is deployed alongside alternating max-pooling layers (Pool-1 to Pool-3) to progressively distill local attribute expressions (e.g., “acoustic technology” and “Bluetooth connectivity”) into highly abstract global feature sets denoted as the heterogeneous semantic matrix U_{TX} . To dynamically adjust feature importance and suppress low-contribution semantic noise, a self-attention layer is integrated over the stacked CNN outputs. The attention weight computation is defined as follows:

$$a_u = \text{softmax}\left(w^\top \tanh(WU_{TX})\right) \quad (17)$$

where U_{TX} denotes the heterogeneous semantic feature matrix, W represents the trainable transformation matrix, and w denotes the attention projection vector. This attention operation adaptively highlights semantically important feature dimensions and suppresses low-contribution semantic noise, thereby improving discriminative semantic representation capability and cross-feature dependency learning.

By aggregating the weighted semantic representations, the model generates a comprehensive heterogeneous feature vector u_d , which captures multi-dimensional user preference characteristics and item attribute semantics for downstream recommendation learning.

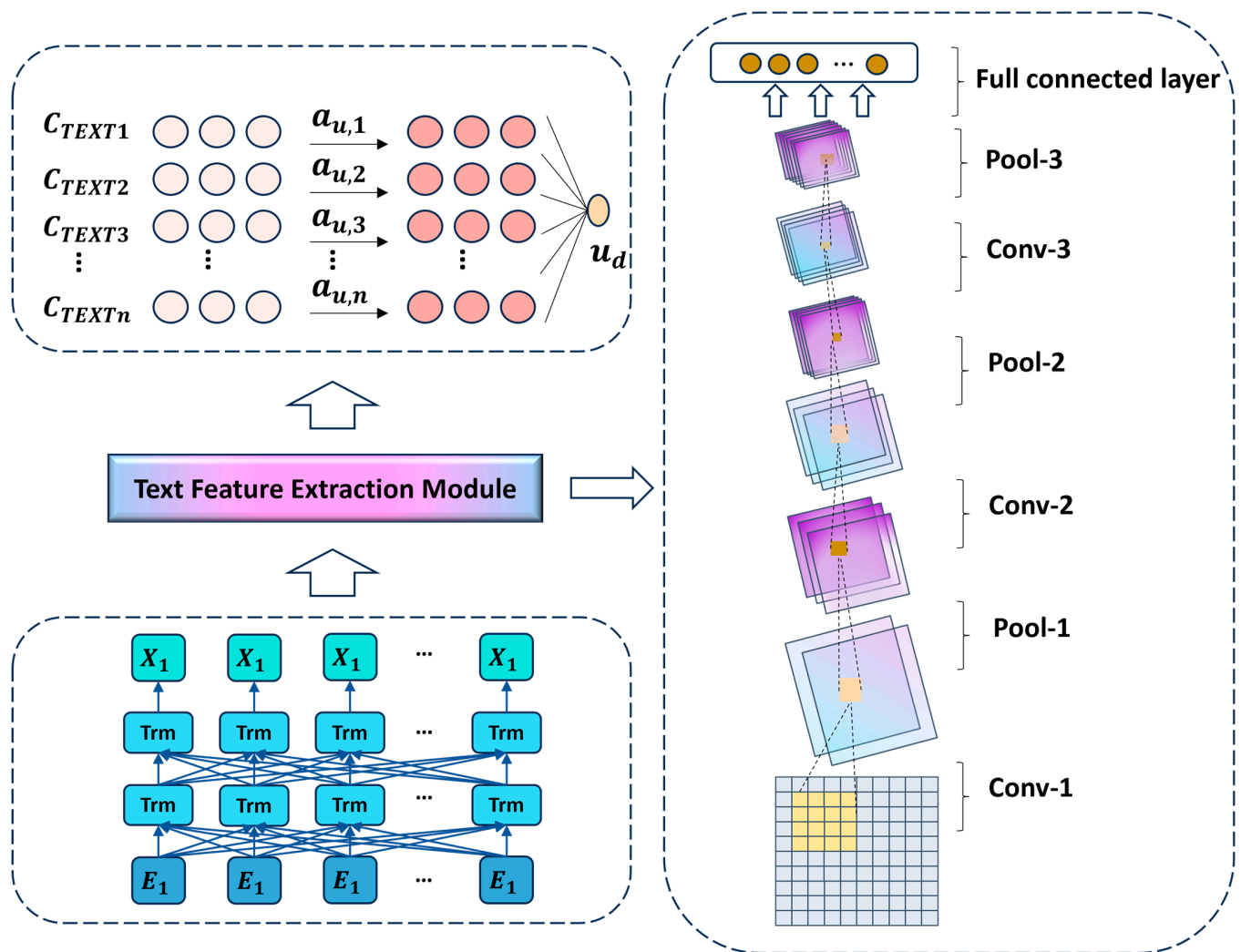


Figure 4. Combining CNN and Self-Attention mechanism text analysis model.

3.4. Multi-Layer Attention-Based Feature Interaction Module

After extracting sentiment, rating, user preference, and product attribute features, effective fusion of multi-source heterogeneous information is required. Traditional methods often fail to capture dynamic feature importance and high-order nonlinear interactions due to fixed weighting and shallow linear fusion. To address these limitations, a multi-layer attention-based feature interaction framework is proposed, consisting of module-level and sub-feature-level attention mechanisms. At the module level, learnable weight matrices combined with Softmax normalization enable global feature importance modeling. At the sub-feature level, a context-aware attention mechanism enhances key dimensions such as sentiment intensity and technical attributes while suppressing noise. Furthermore, a DCN is integrated to explicitly capture high-order feature interactions and jointly exploit implicit semantic relations through a parallel MLP, improving feature fusion and generalization [46].

3.4.1. Module-Level Attention Mechanism

The module-level attention mechanism is designed to enhance model performance by allowing the system to dynamically adjust the contribution of each feature module according to the characteristics of the input data. Feature vectors extracted from each module are transformed into attention scores through a fully connected layer, providing the basis for subsequent feature weighting. The Softmax function is then applied to normalize these scores into effective weight values, representing the relative importance among

different modules. These weights play a crucial role in the subsequent weighted fusion process, ensuring that the recommender system can adaptively adjust the influence of each module in real time under varying contexts.

3.4.2. Sub-Feature Attention Mechanism

The sub-feature attention mechanism is designed to capture fine-grained semantic signals within each feature module. For each module, attention weights are computed over its internal feature dimensions to selectively emphasize informative sub-features while suppressing irrelevant or noisy components. Contextual information is incorporated into the attention computation to enhance adaptability. The refined module representations are obtained via attention-weighted aggregation and subsequently used for downstream recommendation modeling.

3.4.3. Deep & Cross Network

To effectively model explicit high-order feature interactions and enhance model expressiveness, this study employs the DCN-V2 architecture. The module integrates parameterized feature crossing with deep neural representations, achieving complementary optimization between explicit feature combination and implicit relation learning. Compared with conventional cross networks, the DCN improves high-order interaction modeling on sparse data through nonlinear transformation enhancement, residual path optimization, and dual-stream feature fusion.

- Feature Aggregation

Module-level attention is employed to dynamically weight and fuse heterogeneous features, including sentiment signals, implicit ratings, user preferences, and item attributes, capturing their relative contributions to user–item interactions. Meanwhile, sub-feature-level attention refines fine-grained semantic dimensions within each feature type to enhance discriminative signals and suppress noise. The outputs of both attention levels are adaptively aggregated to form a unified feature representation, which is subsequently fed into the cross network for high-order interaction modeling.

- Explicit Feature Crossing

Given the initial feature vector x_0 , the cross network generates high-order cross features through recursive feature transformations. To capture complex, matrix-form structural dependencies efficiently, the cross operation at the l -th layer is mathematically defined based on the DCN-V2 formulation:

$$x_{l+1} = x_0 \odot (W_l x_l + b_l) + x_l \quad (18)$$

where W_l denotes the learnable weight matrix, b_l represents the bias vector, and $\odot$ indicates the Hadamard product. To reduce the number of parameters, cross-layer parameter sharing is applied, where the original feature vector (i.e., the fused feature $\tilde{M}$) serves as the base vector participating in all cross-layer operations. This helps sustain the continuous propagation of both global semantic information and local detailed features throughout the deep network. By constraining the rank of the weight matrix W_l , the parameter complexity is reduced from $O(Ld^2)$ in traditional cross networks to $O(Ld)$, thereby mitigating the risk of overfitting.

In conventional cross networks, linear constraints may lead to incomplete feature interactions. Accordingly, nonlinear enhancement is introduced through the combination of the Hadamard product and linear transformation. Specifically, the Hadamard product enables element-wise nonlinear scaling, while the weight matrix W_l performs parametric

space rotation to enrich feature diversity, allowing the generation of nonlinear feature combination patterns. The output x_l at the l -th layer can thus be interpreted as an $(l + 1)$ -order polynomial expansion of x_0 .

To alleviate gradient vanishing in deep networks, residual connections are incorporated, where each layer retains its input feature x_l via an identity mapping path. This design ensures network stability under highly sparse conditions (e.g., when W_1 is poorly initialized, the network gracefully degrades to an identity function).

- **Implicit Relation Modeling**

To compensate for the limitations of explicit feature crossing in learning complex patterns, a parallel deep neural network is constructed to capture latent feature relationships. The network structure is defined as follows:

$$z_{deep} = \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 x_0 + b_1) + b_2) \quad (19)$$

where $W_1 \in \mathbb{R}^{d_h \times d}$ and $W_2 \in \mathbb{R}^{d_h \times d_h}$ denote the parameters of the fully connected layers, and d_h represents the hidden layer dimension. Through hierarchical nonlinear transformation, the deep network progressively learns latent semantic interaction patterns and implicit preference dependencies embedded within heterogeneous feature representations.

The ReLU activation function introduces non-linear semantic projection capability into the interaction learning process. Crucially, the mathematical zero-gating property of ReLU induces architectural sparsity, which automatically truncates the activation of a vast majority of low-contribution dimensions. This sparse projection mechanism helps extract discriminative interaction vectors while reducing irrelevant semantic background noise in highly sparse recommendation spaces [47]. This nonlinear transformation further improves the representation capability of high-order heterogeneous preference interactions.

In addition, Batch Normalization and Dropout mechanisms are incorporated to improve representation stability, alleviate overfitting risk, and enhance the robustness of heterogeneous semantic learning under sparse recommendation conditions.

- **Explicit–Implicit Feature Fusion**

The explicit cross features and implicit deep features are integrated through the parametric adaptive aggregation mechanism as formulated in Equation (20):

$$z_{interact} = \left[\beta \cdot z_{cross} \parallel (1 - \beta) \cdot z_{deep} \right] \quad (20)$$

where z_{cross} denotes the explicit cross-domain interaction representation, z_{deep} represents the implicit nonlinear semantic representation, and $\beta \in [0, 1]$ is a self-adaptive balancing coefficient dynamically learned via the optimization process. This parametric adaptive aggregation strategy jointly preserves interpretable cross-feature interaction patterns and data-driven latent semantic dependencies, while automatically reconciling the relative contributions of explicit and implicit dual pathways. Consequently, it effectively alleviates manual tuning bias and enhances the completeness and diversity of heterogeneous preference representation learning.

Fundamentally, the explicit cross layers are dedicated to explicitly capturing bounded multi-order polynomial feature interactions, whereas the parallel deep neural network layers specialize in learning implicit, high-order non-linear semantic mappings. Due to this mathematical divergence in their operation paths, the two streams naturally exhibit minimal semantic redundancy and exceptional structural complementarity [34]. Unlike early-stage feature fusion strategies that may suppress complementary semantic information, the proposed fusion mechanism maintains independent semantic propagation paths for explicit and implicit interaction features via the final adaptive aggregation. This design reduces

information interference among heterogeneous feature domains and improves high-order semantic interaction capability during downstream recommendation learning.

3.5. Rating Prediction Module

After feature interaction, the model employs an MLP to map the high-order interaction features into the rating space.

3.5.1. Feature Aggregation and Dimensionality Reduction

The interaction features z_{interact} are connected with the fine-grained feature representations through residual connections to preserve the original semantic information. Subsequently, a fully connected layer is applied to perform feature compression and non-linear transformation:

$$H_1 = \text{ReLU}(W_p H_0 + b_p) \quad (21)$$

where $W_p \in \mathbb{R}^{d_h \times d_c}$ denotes the projection matrix (d_c represents the total dimension of concatenated features, and d_h is the hidden layer dimension), and b_p is the bias term. This operation effectively reduces the feature dimensionality to $\mathbb{R}^{d_h}$, alleviating the curse of dimensionality.

3.5.2. Higher-Order Nonlinear Mapping

Construct a third-order nonlinear transformation layer to progressively approximate the rating function:

$$H_2 = \text{Dropout}(\text{ReLU}(W_1 H_1 + b_1)) \quad (22)$$

$$H_3 = \text{Dropout}(\text{ReLU}(W_2 H_2 + b_2)) \quad (23)$$

$$H_4 = \text{ReLU}(W_3 H_3 + b_3) \quad (24)$$

where $W_i \in \mathbb{R}^{d_h \times d_h}$ denotes the weight matrix and b_i represents the corresponding bias term. The introduction of the Dropout layer enhances the model's generalization ability and mitigates overfitting.

3.5.3. Rating Prediction

The rating prediction layer maps the hidden feature representations to the rating space through a linear transformation:

$$\hat{y} = \sigma(W_o H_4 + b_o) \cdot (y_{\max} - y_{\min}) + y_{\min} \quad (25)$$

Here $W_o \in \mathbb{R}^{1 \times d_h}$ denotes the weight matrix of the output layer, and $\sigma(\cdot)$ represents the Sigmoid activation function, which normalizes the output to the range $[0, 1]$. The parameters $y_{\max} = 5, y_{\min} = 1$ correspond to the rating normalization bounds, ensuring accurate reconstruction within the 1–5 star rating scale.

3.6. Model Optimization Method

To achieve collaborative semantic alignment between emotional representations and recommendation preference interaction patterns, the proposed framework leverages an end-to-end multi-task joint learning strategy. Crucially, this customized regularized optimization routing adaptively balances gradient backpropagation across heterogeneous text, behavior, and structural metadata streams, constituting a pivotal architectural contribution of this study to structurally resolve interaction sparsity and feature heterogeneity.

3.6.1. Sentiment Analysis Training and Optimization

To achieve collaborative optimization between sentiment semantic learning and recommendation preference modeling, the proposed framework adopts a multi-task joint learning strategy in which the sentiment analysis task and the recommendation task share underlying heterogeneous semantic representations and are optimized simultaneously. The joint learning objective enables collaborative semantic alignment between emotional representations and user preference interaction patterns.

The sentiment learning loss (l_s) measures the consistency between predicted emotional semantic representations and observed user feedback signals. The user-side sentiment supervision loss (l_{user}) is defined as follows. For each user u , given the set of historical review documents $D_{u,i}$, the discrepancy between the predicted sentiment score $S_{i,k}$ and the observed user rating $r_{i,k}$ for review $d_{i,k}$ is computed as:

$$l_{user} = \frac{1}{|D_{u,i}|} \sum_{d_{u,k} \in D_{u,i}} (S_{u,k} - r_{u,k})^2 \quad (26)$$

where $S_{u,k}$ denotes the predicted sentiment representation score of the review $d_{u,k}$, and $r_{u,k}$ represents the corresponding observed user rating. This objective function constrains emotional semantic representations to maintain consistency with user preference feedback signals, thereby improving sentiment-aware preference representation learning and enhancing the semantic alignment capability between review sentiment and interaction behavior.

Item-side comment loss (l_{item}):

For each item i , given its associated review set $D_{u,i}$, the mean squared error between the predicted sentiment score $s_{u,k,i}$ and the actual rating $r_{u,k,i}$ for each review $d_{u,k,i}$ is computed as:

$$l_{item} = \frac{1}{|D_{u,i}|} \sum_{d_{u,k,i} \in D_{u,i}} (s_{u,k,i} - r_{u,k,i})^2 \quad (27)$$

This item-side sentiment supervision objective enables the model to learn item-level emotional semantic representations that are consistent with observed user preference feedback patterns.

The final sentiment loss l_s is obtained as the arithmetic mean of the two components, defined as:

$$l_s = \frac{1}{2} (l_{user} + l_{item}) \quad (28)$$

This dual-perspective semantic supervision strategy jointly captures emotional preference dependencies from both user-centric and item-centric semantic spaces, thereby improving heterogeneous sentiment representation learning capability.

The recommendation learning objective is defined as the discrepancy between the predicted preference score $\hat{r}_{u,i}$ and the observed user rating $r_{u,i}$:

$$l_r = (\hat{r}_{u,i} - r_{u,i})^2 \quad (29)$$

This recommendation objective constrains the learned heterogeneous interaction representations to preserve user–item preference consistency during rating prediction learning.

The final joint optimization objective is formulated by dynamically balancing the recommendation learning task and the sentiment semantic learning task through the adjustment coefficient γ , where $0 \leq \gamma \leq 1$:

$$l_{all} = (1 - \gamma)l_r + \gamma l_s \quad (30)$$

When γ approaches 1, the framework emphasizes sentiment semantic representation learning, whereas smaller values of γ place greater emphasis on interaction-based preference modeling. This dynamic balancing strategy enables collaborative optimization between emotional semantic understanding and user–item interaction learning, thereby improving heterogeneous semantic alignment and overall recommendation robustness. To prevent task imbalances and guarantee optimal recommendation accuracy, the selection of this balancing coefficient is empirically evaluated and determined through comprehensive hyperparameter sensitivity analysis (as detailed in Section 4.2.2).

3.6.2. Parameter Update Semantics and Adaptive Optimization Routing

The proposed framework adopts a hybrid shared–specific multi-task architecture. Shared representation layers capture common semantic dependencies across tasks, while task-specific layers preserve independent optimization objectives for sentiment analysis and recommendation prediction.

- (1) **Shared Layer Parameter Updates:** The RoBERTa encoder, topic analysis module, and BiGRU layer act as shared representation layers and receive unified gradients from both tasks. This joint optimization enables the model to learn generalizable semantic representations while aligning user sentiment with product attributes.
- (2) **The sentiment mapping layer and recommendation gating layer are optimized independently using task-specific gradients.** The former predicts sentiment intensity from textual representations, whereas the latter employs a third-order cross network with parallel deep streams to capture high-order user–item interactions. This design alleviates negative transfer between heterogeneous tasks.
- (3) **Adaptive Optimization Routing (AdamW):** AdamW is adopted to optimize all network parameters. By leveraging first- and second-order gradient statistics, it dynamically adjusts parameter updates, alleviates gradient oscillation caused by sparse interactions, and maintains stable optimization for both shared and task-specific layers.
- (4) **Learning Rate Cosine Annealing Schedule:** A Cosine Annealing schedule is incorporated to improve convergence. Larger learning rates in early training encourage global exploration, while gradually decayed learning rates enable fine-grained optimization of latent preference representations and improve overall training stability.

3.6.3. Autoencoder Training and Optimization

The primary objective of the autoencoder training process is to optimize the reconstruction loss so that the model can learn compact latent representations from sparse user–item interaction data, thereby improving rating prediction accuracy.

To address the high sparsity of the rating matrix, gradient updates are restricted to observed (non-zero) entries, reducing computational complexity through sparse matrix representations (e.g., CSR format) and efficient sparse–dense matrix operations. To prevent overfitting, an L2 regularization term is added to the loss function, constraining the Frobenius norm of the encoder and decoder weight matrices, with regularization strength controlled by hyperparameter λ .

The ReLU activation function is employed in the encoder to enhance nonlinearity, while Sigmoid is used in the decoder to constrain the reconstruction range. Mini-batch stochastic gradient descent combined with the Adam optimizer accelerates convergence. Batch normalization mitigates internal covariate shift, and gradient clipping stabilizes training by preventing exploding gradients. These optimizations collectively help achieve efficient convergence and generalization across sparse rating scenarios.

3.6.4. Text Feature Analysis Model Training and Optimization

The text analysis model is optimized through a two-stage progressive fine-tuning strategy.

In the first stage (local pretraining), all parameters of the RoBERTa encoder are frozen to preserve semantic consistency, while CNN and attention layers are trained using the AdamW optimizer with dropout and attention regularization. Early stopping is triggered when validation loss stagnates for three consecutive epochs.

In the second stage (end-to-end fine-tuning), the last four Transformer layers of RoBERTa are unfrozen, and layer-wise learning rates are adopted (1×10^{-5} for RoBERTa, 3×10^{-4} for CNN and attention modules). A cosine annealing schedule gradually decays the learning rate to 1×10^{-6} over 20 epochs. The training objective employs a weighted MSE loss to emphasize high-rating samples and reduce long-tail bias, while gradient clipping and exponential moving average (EMA) are applied to help boost stability and convergence efficiency.

3.6.5. Feature Interaction Module Training and Optimization

The multi-level attention-based feature interaction module adopts a three-stage hierarchical training paradigm to capture cross-domain dependencies.

First, in the module-level attention stage, global importance among sentiment, rating, user preference, and product attributes is dynamically estimated via learnable weight matrices and Softmax normalization, producing a 256-dimensional global interaction vector.

Next, in the sub-feature enhancement stage, a context-aware attention unit refines key semantic dimensions, while feature-level dropout randomly masks 20% of dimensions to mitigate sparsity and noise.

Finally, in the deep cross-network stage, explicit feature crosses model high-order user-item interactions, while parallel MLPs capture implicit semantic relations. The outputs of both paths are fused via adaptive feature aggregation.

The model incorporates progressive parameter unfreezing, mixed-precision training, and Gaussian noise injection to improve robustness and convergence. Gradient clipping and early stopping (triggered after 10 stagnant epochs) help stabilize training and prevent overfitting.

4. Experiments and Results Analysis

4.1. Dataset Selection and Evaluation Method

4.1.1. Dataset

In the Section 4, this study employs the publicly available Amazon Product Dataset [48] to evaluate the proposed model. The dataset incorporates a three-level information structure: User interaction level: includes rating data, multimodal review content, and helpfulness votes; Product metadata level: integrates product descriptions, category tags, price history, brand information, and image features; Product relationship network: establishes implicit relational graphs through “users who bought this item also bought” and “similar items” links. To verify the model’s effectiveness and adaptability, five representative subcategories were selected: Movies and TV, Digital Music, Musical Instruments, Toys and Games, and Home and Kitchen. To mitigate data extreme sparsity and noise, this study directly adopts the officially provided 5-core subsets, in which all unhandled raw interactions have already been iteratively filtered by the dataset creators so that every retained user and item has at least five interaction records.

Let N_U , N_I , and N_R denote the total numbers of users, items, and observed user–item interaction records, respectively. Data sparsity measures the proportion of unobserved interactions within the user–item matrix, formally calculated as:

$$Sparsity = 1 - N_R / (N_U \times N_I) \quad (31)$$

The selected subsets span diverse domains and vary significantly in dataset scale and sparsity, enabling a comprehensive evaluation of the model’s performance across different application contexts. The detailed statistical characteristics of these five 5-core datasets are summarized in Table 2, where the statistics correspond directly to the symbols defined above and the sparsity values are calculated according to Equation (31).

Table 2. Detailed statistical information of the datasets.

Dataset	Number of Users (N_U)	Number of Items (N_I)	Number of Interactions (N_R)	Sparsity (%)
Movies and TV	123,960	50,052	1,697,533	99.97
Digital Music	5426	3568	64,475	99.67
Musical Instruments	1397	900	10,261	99.19
Toys and Games	19,412	11,924	167,597	99.93
Home and Kitchen	66,519	28,237	551,682	99.97

4.1.2. Dataset Processing

This study employs the publicly available Amazon dataset as the primary data source. The dataset stores a massive number of user purchase records in JSON format (as illustrated in Figure 5). Each JSON entry contains several key fields: user identifiers (*reviewer ID* and *reviewer name*), item identifiers (*asin*), interaction content (*review Text* for full review text, *summary* for review title, and *overall* for 1–5 star ratings), and temporal information (*Unix Review Time* as a timestamp and *review Date* as a human-readable date).

```
{
  "reviewerID": "B3J7K5Y8U9I2R",
  "asin": "1345363274",
  "reviewerName": "David Lee \"Just loving it!\",",
  "helpful": [0, 0],
  "reviewText": "This product is a game-changer. It's so easy to use
and has really improved my daily workflow.",
  "overall": 4.5,
  "summary": "Excellent product",
  "unixReviewTime": "1325678900",
  "reviewTime": "11 15, 2012"
}
```

Figure 5. Json file storage format.

A systematic text preprocessing procedure was applied to the Amazon user review data, aiming to remove redundant symbols (such as nonstandard punctuation and HTML tags) and low-frequency noise words caused by user-generated content. The process includes the following steps: first, the JSON-formatted purchase records were parsed into JSON objects, extracting key fields such as *reviewText* and storing them in a string list. Next, text cleaning was performed by converting all text to lowercase using `toLowerCase()`, and regular expressions were employed to remove extra spaces, abnormal symbols, and rare words with a frequency of less than five. The cleaned and normalized text was then

saved as UTF-8 encoded files for subsequent analysis. Figure 6 illustrates a comparison between the raw and cleaned text.

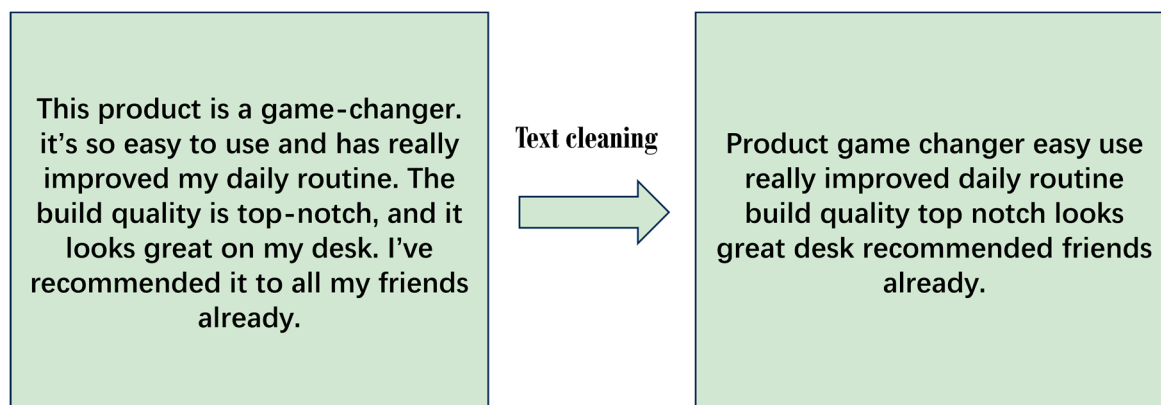


Figure 6. Text cleaning example.

Each review is assigned a unique identifier (ID) to facilitate fast data retrieval for the model, thereby ensuring precise mapping between model inputs and their corresponding review texts. To boost computational efficiency, an ID-based index is constructed for all reviews to substantially optimize data storage and access speed. Notably, all raw user and item IDs are uniformly converted into integers before dataset construction to satisfy standardized preprocessing requirements, as visualized in Figure 7.

ID	User identifier	ID	Product Identifier
2345	B5T2U7KLP8RXYV	347	C002M09FKS
5678	D3R6F9H4W2QZAS	789	C006N12GHB
1236	E7Y8I9O5P4LJMX	456	C000Z07RTE
9870	F2G3J6K7LIVUXW	876	C001A15YHB
4563	H9I8U7Y6T5R4E3	234	C003D08FKG
7891	J1K2L3M4N5O6P	567	C005B10LJH
3452	Q9W8E7R6T5Y4U3	123	C007Z14HKB
6780	R2T3Y4U5I6O7P	987	C008F16SXD
8901	S3D4F5G6H7J8K	654	C009G17VNB
1023	T6Y7U8I9O0P1Q	321	C004C11RFP
2134	U9I8O7P6L5K4J	432	C000X03LHG
3245	V2B3N4M5K6J7H	765	C001W13DMC
4356	W7R6E5T4Y3U2I	876	C002Q09KLP

Figure 7. Mapping relationship between users and products.

Each purchase interaction in this study was uniquely identified using a tuple of (user ID, item ID), which was mapped to the corresponding raw data to enable efficient information retrieval. The dataset is partitioned into training, validation, and testing sets in a 7:2:1 ratio. To ensure a strictly fair comparison, all baseline models and the proposed HAIRec model are trained, validated, and evaluated on the exact same data splits, eliminating any potential data-side bias. To maintain the basic connectivity of the collaborative filtering graph, each user/item is ensured to have at least one interaction in the training set. This partition methodology strictly follows the well-established evaluation protocol in review-based recommendation literature (e.g., DeepCoNN [21], UniDNR [49]).

Under this standard protocol, the challenge shifts to evaluating “weak cold-start” and “extreme interaction sparsity,” as more than 74% of users possess fewer than 3 behavioral footprints in the training phase.

To strictly prevent target task leakage, it is guaranteed that when predicting the rating $r_{u,i}$ for a target interaction (u, i) , the review text $d_{u,i}$ associated with this specific rating is strictly excluded from the multi-source feature learning module. Instead, the model only utilizes the historical reviews written by the user u on other items (excluding i) to construct the user profile, and the reviews written by other users on item i (excluding u) to construct the item profile. Here, “historical reviews” in this static rating prediction context refer to all available review documents in the dataset other than the review associated with the target user–item interaction. This leave-target-out strategy ensures that no review text or sentiment information from the target interaction is leaked into the feature representation learning process.

Since the objective of this study is static rating prediction rather than future interaction forecasting, the dataset is partitioned using a random splitting strategy instead of chronological splitting. To ensure sufficient training coverage, each user and each item are guaranteed to have at least one associated interaction included in the training set before the remaining samples are randomly divided into the training, validation, and testing sets.

Specifically, the dataset was represented as a set of identifiers E , which was grouped into two mappings: user2item and item2user. During the construction of the training set, each user in the user2item mapping was first ensured to have at least one associated item included. Next, the item2user mapping was traversed to include interaction records for items not yet covered. The remaining unassigned data were then used to fill the training set to 70% of the total dataset size. The remaining samples were distributed into the validation set (20%) and testing set (10%) accordingly.

4.1.3. Evaluation Method

To ensure a fair and comprehensive performance comparison between the proposed framework and baseline recommendation models under identical evaluation settings, three widely adopted evaluation metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), are employed for rating prediction performance assessment.

RMSE is calculated as the square root of MSE and is more sensitive to large prediction deviations, while MAE reflects the average absolute discrepancy between predicted ratings and actual ratings. Lower values of MSE, RMSE, and MAE indicate stronger consistency between predicted preference representations and observed user feedback, reflecting improved recommendation accuracy and preference modeling capability.

To improve evaluation reliability and reduce the influence of stochastic factors, all experiments were independently repeated five times under the same data partition settings with different random seeds. In each run, the random seed affected parameter initialization, mini-batch sampling order, dropout operations, and other stochastic optimization processes. The final performance results were obtained by averaging the experimental outcomes, and the corresponding standard deviations were additionally reported.

Furthermore, paired two-tailed t-tests with Holm–Bonferroni correction were conducted between HAIRec and baseline methods to verify the statistical significance of the observed performance improvements.

4.2. Experimental Environment and Parameter Settings

4.2.1. Experimental Environment

The experiments were conducted on a Windows 11 platform using Python 3.9 as the primary programming language. The proposed model was implemented with the PyTorch deep learning framework. The hardware environment comprised an Intel Core Ultra 5 125H CPU and an NVIDIA GeForce RTX 4060 GPU, utilizing CUDA and cuDNN for efficient parallel acceleration.

4.2.2. Parameter Settings

Both sentiment and recommendation tasks are optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-5} . Gradient clipping with a threshold of 5.0 and a cosine annealing learning rate scheduler with warm-up are applied. The training batch size is 256, and early stopping is triggered if the validation loss does not improve for 10 consecutive epochs.

For the autoencoder, the Adam optimizer is adopted with a learning rate of 1×10^{-3} , weight decay of 1×10^{-4} , and L2 regularization weight $\lambda = 0.01$. During the training phase, the RoBERTa backbone network remains frozen, while only the light CNN and attention layers are updated with a learning rate of 3×10^{-4} and a dropout rate of 0.2.

The explicit path of the DCN-V2 is configured with 3 cross layers, while the implicit path comprises three fully connected layers with dimensions of 768, 512, and 256, respectively. The sensitivity analysis of the task balance coefficient γ was conducted by systematically evaluating candidate values in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Due to space limitations, the complete quantitative results on the representative Movies and TV dataset are reported in Table 3. As observed, $\gamma = 0.5$ achieves the optimal performance across MSE, RMSE, and MAE, and is therefore selected as the default setting (with identical performance trends consistently observed across all other benchmark datasets). Additionally, the LDA topic count is set to $K = 20$. For all baseline models, their optimal hyperparameters—specifically the regularization coefficient $\lambda \in \{0.001, 0.01, 0.1\}$ and latent dimension $d \in \{32, 64, 128\}$ —were rigorously determined via grid search on the validation sets to help realize a fair comparison.

Table 3. Sensitivity analysis of the task balance coefficient γ on the Movies and TV dataset.

γ	MSE (Mean $\pm$ SD)	RMSE (Mean $\pm$ SD)	MAE (Mean $\pm$ SD)
0.1	0.923 $\pm$ 0.006	0.961 $\pm$ 0.005	0.896 $\pm$ 0.004
0.3	0.899 $\pm$ 0.005	0.948 $\pm$ 0.004	0.887 $\pm$ 0.003
0.5	0.884 $\pm$ 0.005	0.940 $\pm$ 0.004	0.882 $\pm$ 0.003
0.7	0.892 $\pm$ 0.005	0.944 $\pm$ 0.004	0.885 $\pm$ 0.003
0.9	0.913 $\pm$ 0.006	0.956 $\pm$ 0.005	0.891 $\pm$ 0.004

4.3. Experimental Results and Analysis

To comprehensively evaluate the performance and robustness of the proposed model, comparisons were conducted against seven representative baseline methods, encompassing both classical review-aware architectures and recent neural recommendation paradigms: PMF, HFT, ConvMF, DeepCoNN, D-Attn, NARRE, and the graph-based benchmark LightGCN. Three widely adopted evaluation metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), were employed across five real-world datasets. The statistical significance of performance improvements was further validated using paired two-tailed *t*-tests with Holm–Bonferroni correction.

Table 4 reports the complete numerical results of all methods across all datasets, including mean values and standard deviations. Figure 8 further visualizes the overall performance trends of different recommendation models.

Table 4. Performance comparison of all methods across five datasets (Mean $\pm$ SD).

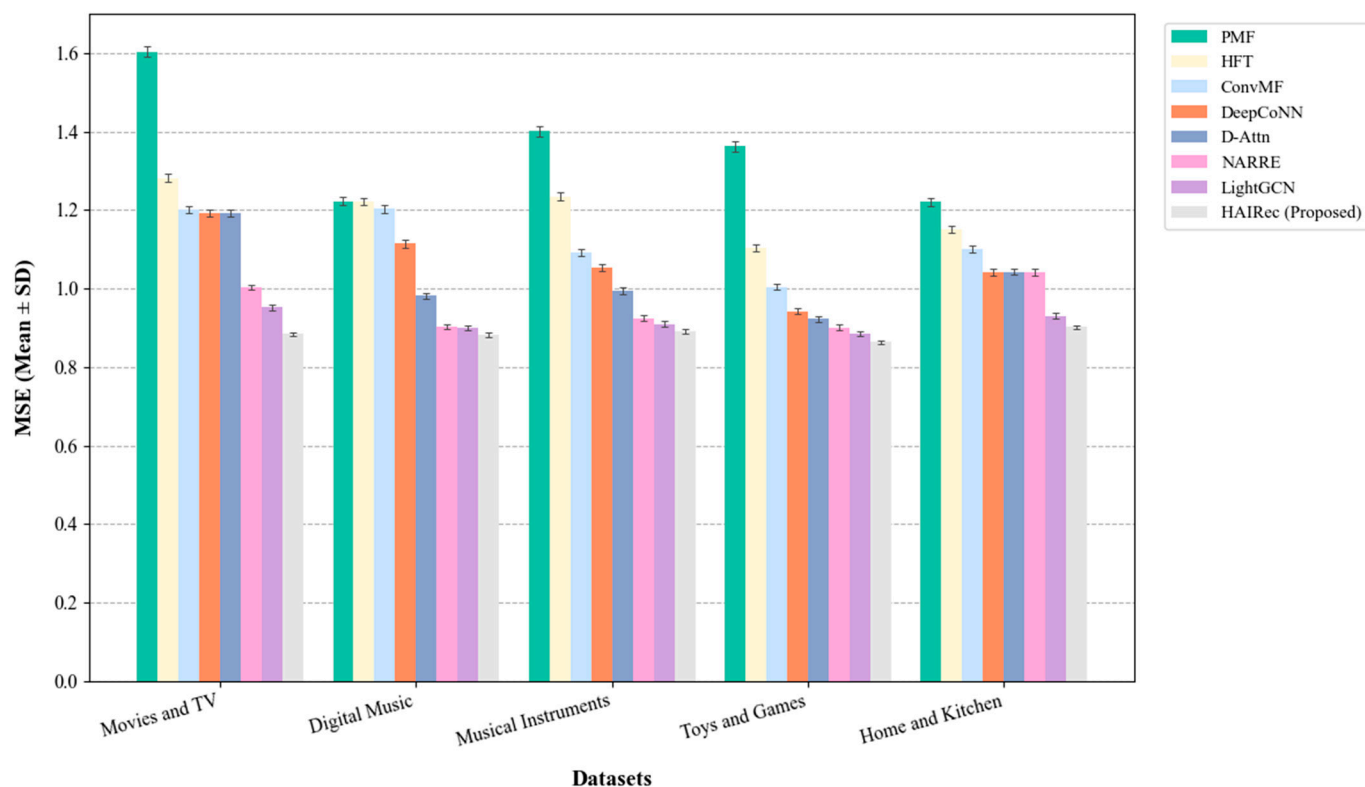
Dataset	Metric	PMF	HFT	ConvMF	DeepCoNN	D-Attn	NARRE	LightGCN	HAIRec (Proposed)
Movies & TV	MSE	1.604 $\pm$ 0.014	1.282 $\pm$ 0.011	1.201 $\pm$ 0.009	1.192 $\pm$ 0.008	1.191 $\pm$ 0.009	1.003 $\pm$ 0.007	0.952 $\pm$ 0.008	0.884 $\pm$ 0.005 *
	RMSE	1.266 $\pm$ 0.009	1.132 $\pm$ 0.008	1.096 $\pm$ 0.007	1.092 $\pm$ 0.006	1.091 $\pm$ 0.006	1.001 $\pm$ 0.005	0.976 $\pm$ 0.006	0.940 $\pm$ 0.004 *
	MAE	1.018 $\pm$ 0.007	0.951 $\pm$ 0.006	0.922 $\pm$ 0.005	0.927 $\pm$ 0.004	0.902 $\pm$ 0.004	0.898 $\pm$ 0.004	0.892 $\pm$ 0.005	0.882 $\pm$ 0.003 *
Digital Music	MSE	1.223 $\pm$ 0.011	1.221 $\pm$ 0.009	1.203 $\pm$ 0.010	1.114 $\pm$ 0.009	0.982 $\pm$ 0.007	0.902 $\pm$ 0.006	0.899 $\pm$ 0.006	0.882 $\pm$ 0.005 *
	RMSE	1.106 $\pm$ 0.008	1.105 $\pm$ 0.007	1.097 $\pm$ 0.007	1.055 $\pm$ 0.006	0.991 $\pm$ 0.005	0.950 $\pm$ 0.005	0.948 $\pm$ 0.005	0.939 $\pm$ 0.004 *
	MAE	0.942 $\pm$ 0.005	0.928 $\pm$ 0.004	0.903 $\pm$ 0.005	0.897 $\pm$ 0.004	0.893 $\pm$ 0.004	0.887 $\pm$ 0.003	0.884 $\pm$ 0.004	0.879 $\pm$ 0.003 *
Musical Instruments	MSE	1.401 $\pm$ 0.013	1.234 $\pm$ 0.010	1.092 $\pm$ 0.008	1.053 $\pm$ 0.008	0.994 $\pm$ 0.008	0.924 $\pm$ 0.007	0.910 $\pm$ 0.007	0.891 $\pm$ 0.005 *
	RMSE	1.184 $\pm$ 0.009	1.112 $\pm$ 0.007	1.045 $\pm$ 0.006	1.026 $\pm$ 0.005	0.997 $\pm$ 0.006	0.961 $\pm$ 0.005	0.954 $\pm$ 0.005	0.944 $\pm$ 0.004 *
	MAE	1.046 $\pm$ 0.007	0.967 $\pm$ 0.005	0.939 $\pm$ 0.004	0.921 $\pm$ 0.005	0.916 $\pm$ 0.004	0.908 $\pm$ 0.004	0.895 $\pm$ 0.004	0.891 $\pm$ 0.003 *
Toys & Games	MSE	1.362 $\pm$ 0.012	1.103 $\pm$ 0.009	1.004 $\pm$ 0.007	0.942 $\pm$ 0.007	0.923 $\pm$ 0.007	0.901 $\pm$ 0.007	0.885 $\pm$ 0.006	0.863 $\pm$ 0.004 *
	RMSE	1.167 $\pm$ 0.008	1.050 $\pm$ 0.006	1.002 $\pm$ 0.005	0.971 $\pm$ 0.005	0.961 $\pm$ 0.005	0.949 $\pm$ 0.005	0.941 $\pm$ 0.005	0.929 $\pm$ 0.003 *
	MAE	0.961 $\pm$ 0.006	0.912 $\pm$ 0.005	0.885 $\pm$ 0.004	0.878 $\pm$ 0.004	0.870 $\pm$ 0.004	0.865 $\pm$ 0.003	0.861 $\pm$ 0.004	0.857 $\pm$ 0.003 *
Home and Kitchen	MSE	1.221 $\pm$ 0.010	1.152 $\pm$ 0.009	1.101 $\pm$ 0.009	1.041 $\pm$ 0.008	1.043 $\pm$ 0.008	1.042 $\pm$ 0.008	0.930 $\pm$ 0.007	0.902 $\pm$ 0.005 *
	RMSE	1.105 $\pm$ 0.007	1.073 $\pm$ 0.007	1.049 $\pm$ 0.006	1.020 $\pm$ 0.006	1.021 $\pm$ 0.006	1.021 $\pm$ 0.005	0.964 $\pm$ 0.005	0.950 $\pm$ 0.004 *
	MAE	1.030 $\pm$ 0.006	0.958 $\pm$ 0.005	0.927 $\pm$ 0.004	0.933 $\pm$ 0.004	0.915 $\pm$ 0.004	0.911 $\pm$ 0.003	0.897 $\pm$ 0.004	0.890 $\pm$ 0.003 *

Note: The best performance values are highlighted in bold. Standard deviations (SD) are calculated over five independent runs. The symbol * indicates that the improvement of HAIRec over all baseline methods is statistically significant at the adjusted significance level ($p < 0.05$) after applying the Holm–Bonferroni correction.

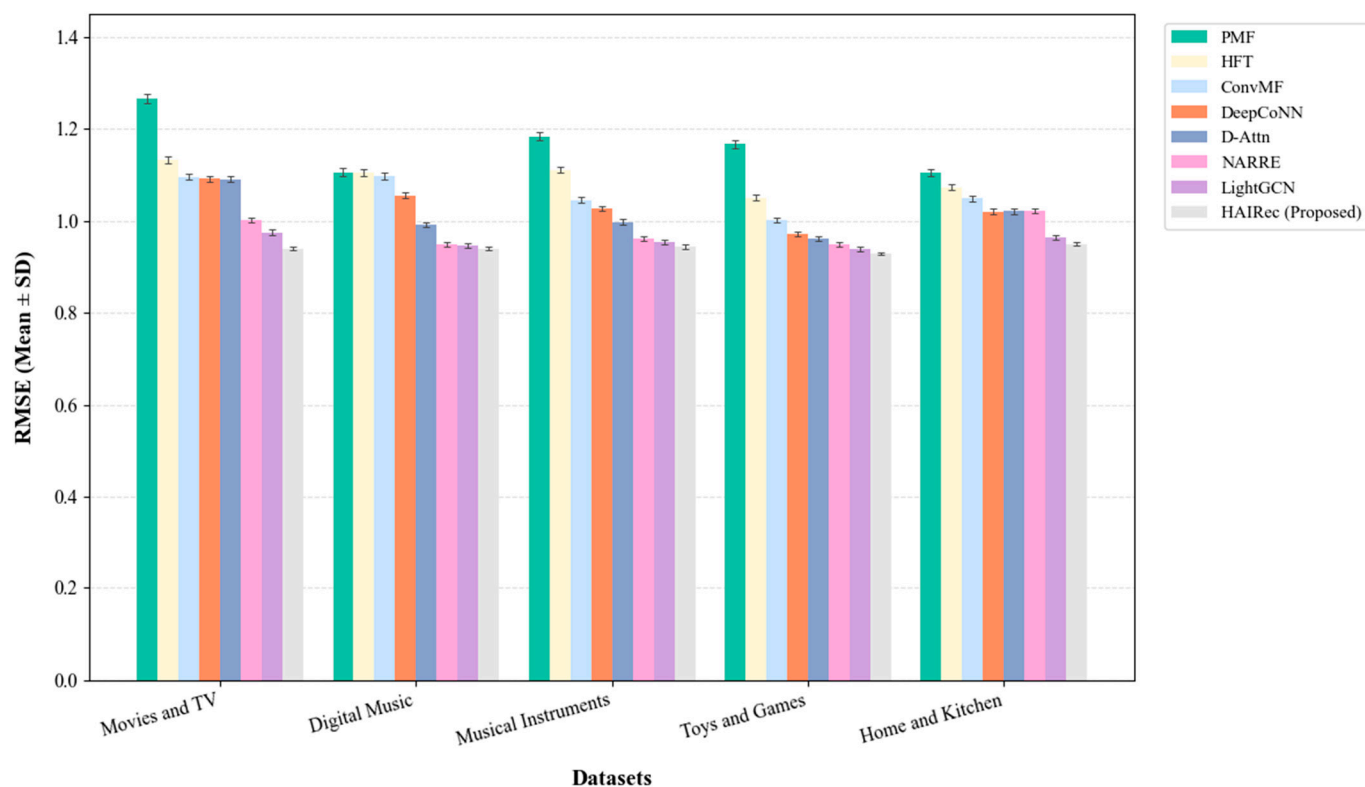
Figure 8 compares multiple recommendation models on five real-world datasets via MSE, RMSE and MAE. The proposed HAIRec achieves the lowest prediction errors against all baselines including PMF, HFT, ConvMF, DeepCoNN, D-Attn, NARRE and the graph-based SOTA LightGCN, showing stronger generalization and more precise user preference modeling.

Although LightGCN effectively captures high-order collaborative filtering signals through graph convolution, it ignores contextual sentiment semantics in user reviews. Instead, HAIRec adopts a dual-awareness paradigm to jointly model user-item structural interactions and review-based sentiment information, thus consistently outperforming all deep learning competitors. The complete numerical results presented in Table 4 further confirm that HAIRec establishes consistent performance advantages across all datasets

and evaluation metrics by successfully modeling heterogeneous semantic dependencies between user-item interactions and review sentiments.



(a)



(b)

Figure 8. Cont.

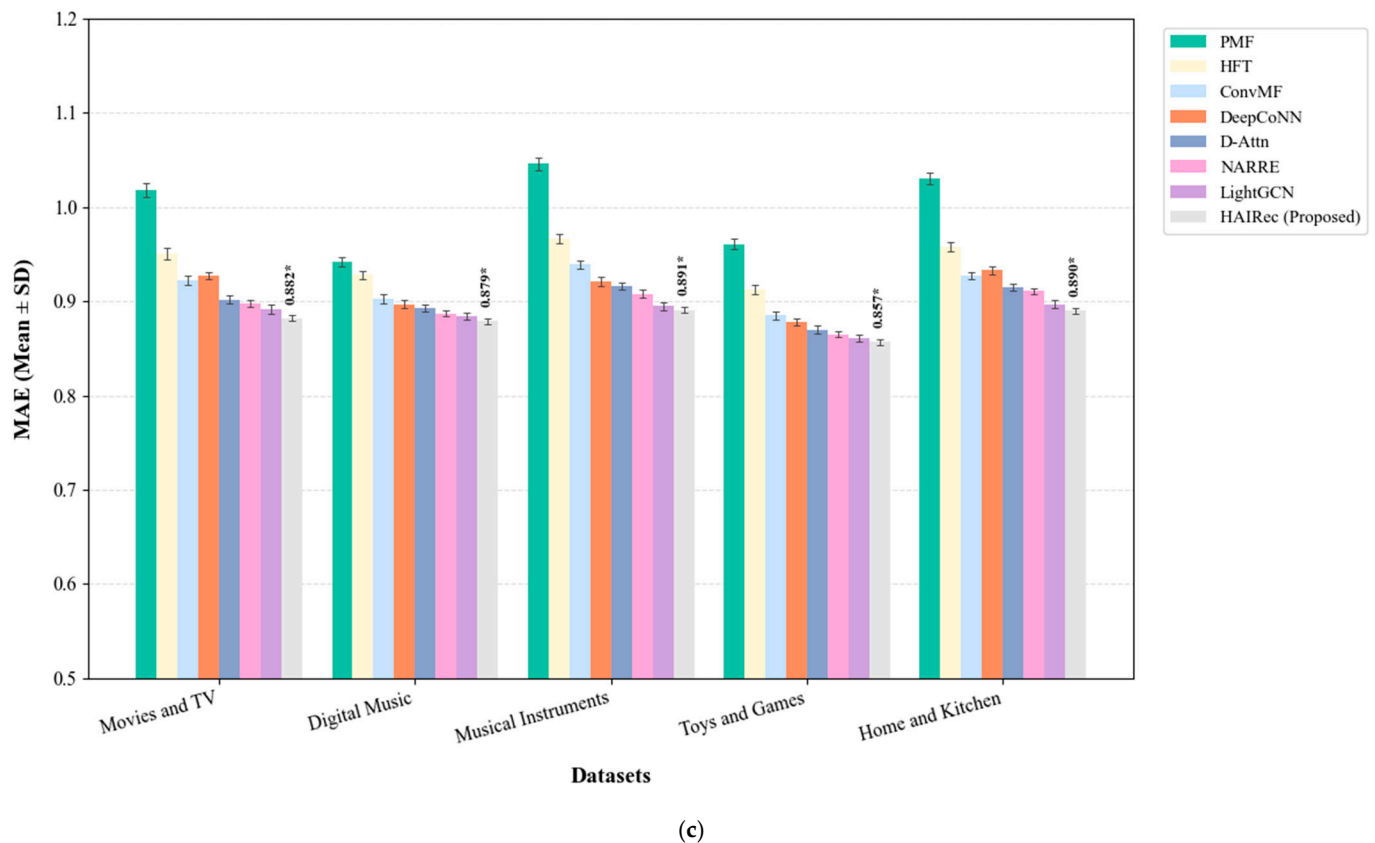


Figure 8. Comparison results of different models under (a) MSE, (b) RMSE, and (c) MAE metrics. The symbol * indicates that the performance difference between HAIRec (Proposed) and the corresponding baseline model is statistically significant at ($p < 0.05$).

HAIRec delivers prominent performance on Movies and TV, Toys and Games, and Home and Kitchen datasets, where user preferences are closely linked to objective item attributes. For Digital Music and Musical Instruments datasets characterized by subjective emotional content, sentiment-aware representation learning enables the proposed HAIRec framework to maintain stable predictive performance.

Traditional matrix factorization methods such as PMF suffer severely from data sparsity, while text-aware deep models like DeepCoNN and NARRE fail to build sufficient heterogeneous feature interactions and adaptive fusion. Equipped with module/sub-feature attention and a deep cross network, HAIRec gains stronger feature representation and prediction accuracy.

The standard deviations observed across repeated runs remain relatively low. This stable optimization across random seeds is primarily attributed to two architectural factors: (1) the usage of normalized, static text representations extracted from pre-trained language backbones, which provides consistent semantic representations and stabilizes early gradient updating trajectories compared to training embeddings from scratch; and (2) the extensive integration of Layer Normalization and residual connections throughout the Cross Network and attention layers, which effectively prevents severe variance during parameter re-initialization.

To rigorously verify the statistical significance of performance gains under these stable runs, the Holm–Bonferroni correction was applied to all 105 pairwise comparisons between HAIRec and baseline models across five datasets and three evaluation metrics ($7 \text{ models} \times 5 \text{ datasets} \times 3 \text{ metrics} = 105$). The complete matrix of adjusted p -values is provided in Table A1 in Appendix A. All adjusted p -values remain well below 0.05, confirming statistical significance. Furthermore, the paired t -statistics ($t(4) \in [4.32, 18.65]$)

and Cohen's d ($d \in [1.93, 8.34]$) confirm that the observed gains correspond to large and meaningful effect sizes.

In summary, this sentiment-interaction hybrid HAIRec model enhances user preference modeling and rating prediction, with reliable feasibility, stability and generalization across diverse datasets.

4.4. Ablation Studies and Structural Analysis

To systematically evaluate the empirical effectiveness of each feature extraction module, the hierarchical multi-level attention mechanism, and the error-driven regularization components, a total of eight fine-grained variant models were designed and analyzed. To provide structural clarity, the specific architectural configurations of the complete HAIRec model and all its corresponding ablation benchmarks are explicitly defined via the configuration matrix in Table 5 and summarized as follows:

Table 5. Architectural Configurations of HAIRec and Its Ablation Variants.

Variant Name	Sentiment Extraction	Rating Latent	User Preference	Product Attribute	Module-Level Attention	Sub-Feature Attention	Dynamic Masking
HAIRec (Full)	✓	✓	✓	✓	✓	✓	✓
HAIRec-w/o-SE	×	✓	✓	✓	✓	✓	✓
HAIRec-w/o-RL	✓	×	✓	✓	✓	✓	✓
HAIRec-w/o-UP	✓	✓	×	✓	✓	✓	✓
HAIRec-w/o-PA	✓	✓	✓	×	✓	✓	✓
HAIRec-w/o-ModAtt	✓	✓	✓	✓	×	✓	✓
HAIRec-w/o-SubAtt	✓	✓	✓	✓	✓	×	✓
HAIRec-w/o-Att	✓	✓	✓	✓	×	×	×
HAIRec-w/o-DM	✓	✓	✓	✓	✓	✓	×

- (1) HAIRec-w/o-SE: Removes the review-based sentiment feature extraction module entirely while retaining rating latent, user preference, and product attribute features for interaction modeling.
- (2) HAIRec-w/o-RL: Removes the rating latent feature extraction module (the implicit collaborative filtering stream) while preserving all remaining behavioral and semantic feature fields.
- (3) HAIRec-w/o-UP: Removes the user numerical preference feature extraction module to isolate the impact of explicit behavioral statistics.
- (4) HAIRec-w/o-PA: Removes the text-encoded product attribute feature extraction module from the prediction pipeline.
- (5) HAIRec-w/o-ModAtt: The module-level attention mechanism is deactivated, retaining only the sub-feature-level attention to evaluate the standalone contribution of fine-grained intra-domain feature weighting.
- (6) HAIRec-w/o-SubAtt: The sub-feature-level attention mechanism is deactivated, preserving only the module-level attention to assess the baseline utility of global inter-domain importance modeling.
- (7) HAIRec-w/o-Att: Removes the entire hierarchical multi-level attention-based feature fusion module, substituting it with a standard three-layer MLP for feature cross-modeling.

- (8) HAIRec-w/o-DM: Retains the complete hierarchical attention architecture but entirely deactivates the error-driven dynamic token masking regularizer (the error-inducing token sets Q_p and Q_n are bypassed) to isolate the standalone utility of proactive token suppression gates.

4.4.1. Evaluation of the Hierarchical Attention Architecture

To investigate how the depth and hierarchy of attention layers influence model performance, a comparative evaluation was executed across all five datasets. Figure 9 presents the ablation results tracking the Mean Squared Error (MSE) metrics of the complete model against HAIRec-w/o-ModAtt and HAIRec-w/o-SubAtt.

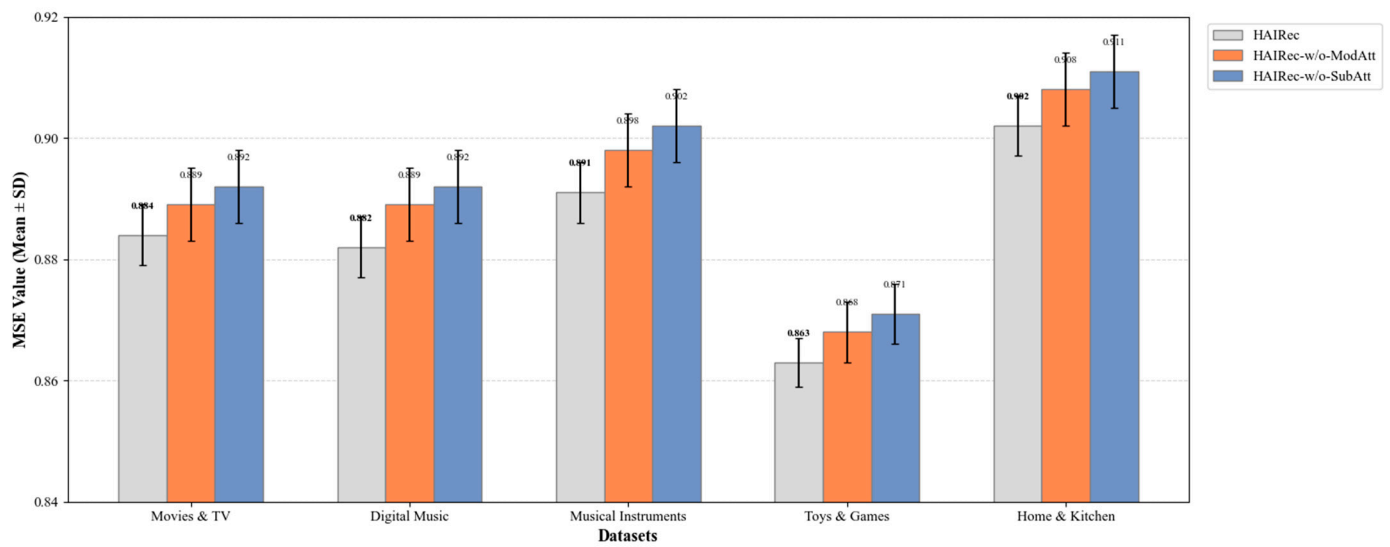


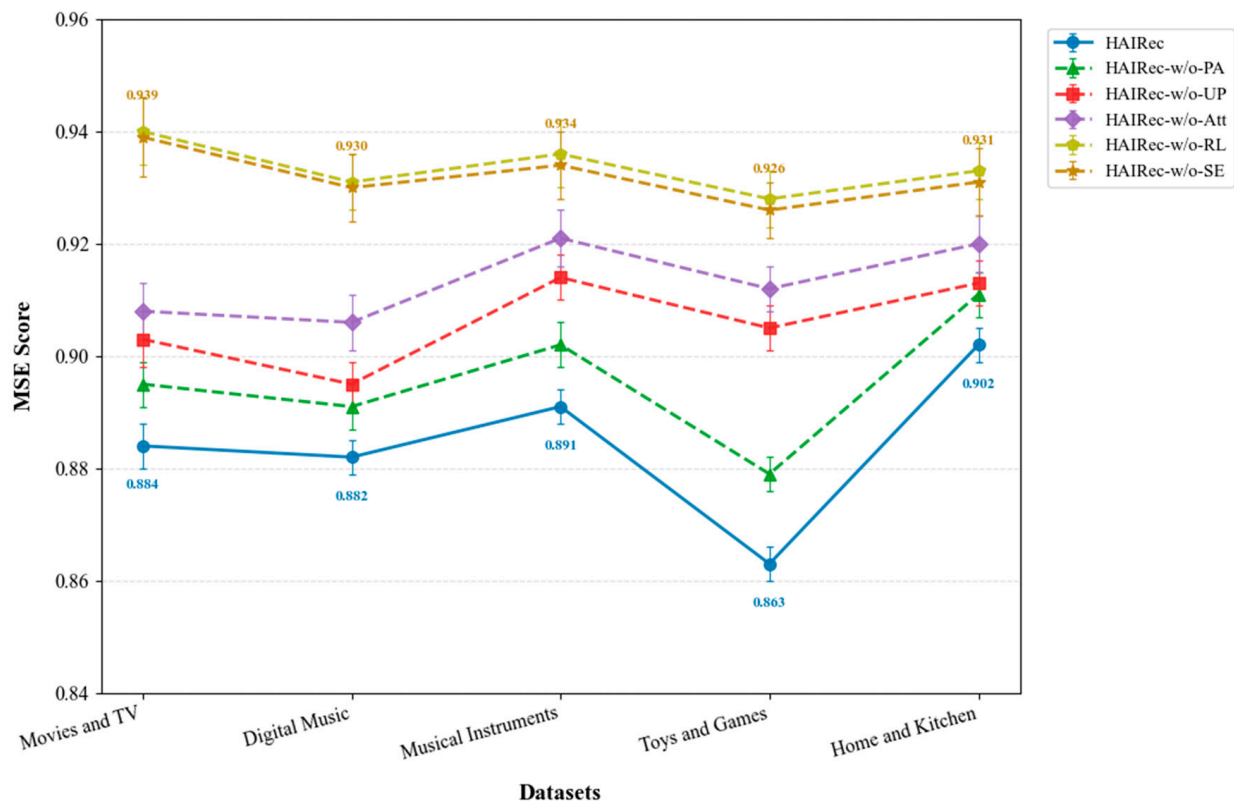
Figure 9. Evaluation and component analysis of the hierarchical attention architecture.

As observed, both simplified variants consistently underperform compared to the complete HAIRec model across all testing environments. Stripping either attention layer yields a noticeable drop in recommendation accuracy. This performance gap demonstrates that the dual-level attention synergy is essential for cohesive feature routing. The experimental results verify that without the joint modeling of macroscopic cross-domain weights (evaluated by removing module-level attention) and local semantic alignments (evaluated by removing sub-feature attention), the model becomes highly sensitive to cross-domain noise, thereby confirming the complementary contributions of both layers to prediction accuracy and stability.

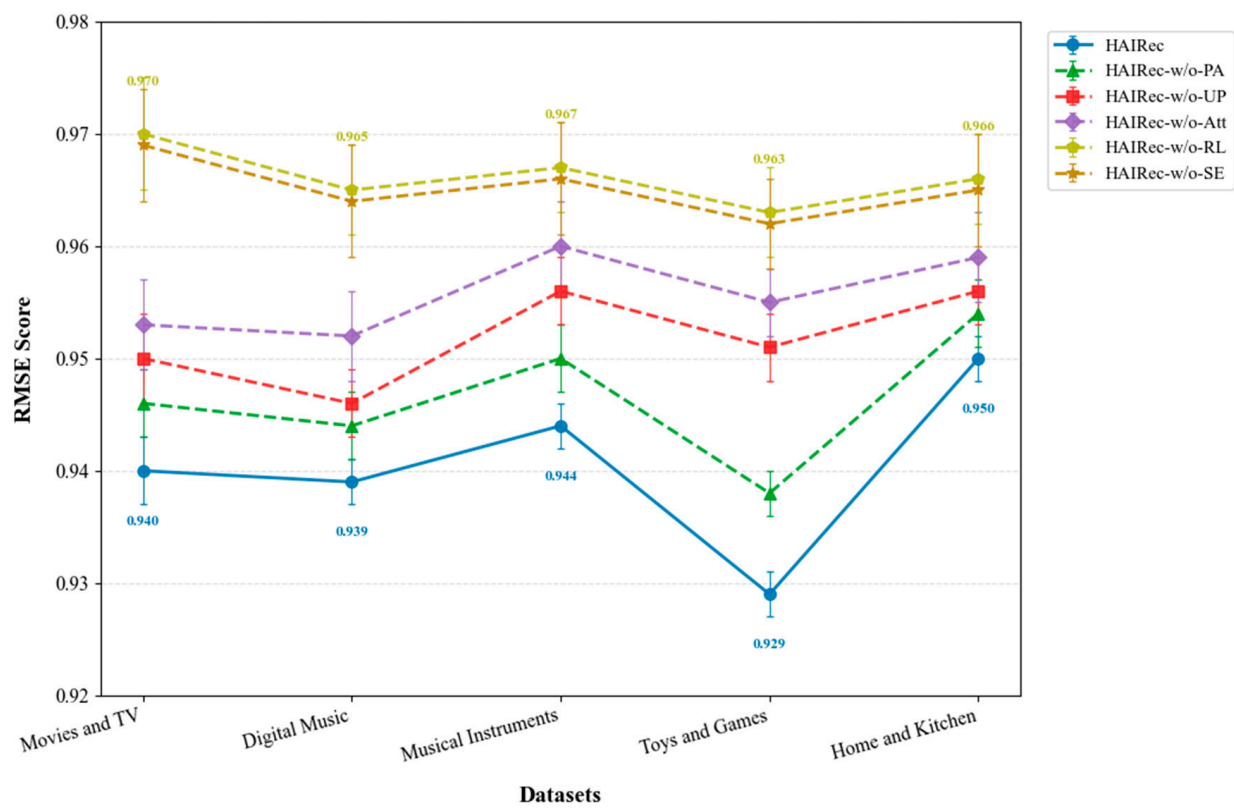
4.4.2. Overall System-Level Component Ablation

The comparative performance results of these five variants and the full HAIRec model across all three metrics (MSE, RMSE, and MAE) and five datasets are visually illustrated in Figure 10a–c.

The experimental results across Figure 10a–c indicate that the proposed HAIRec model consistently outperforms its ablated versions—HAIRec-w/o-SE, HAIRec-w/o-RL, HAIRec-w/o-UP, HAIRec-w/o-PA, and HAIRec-w/o-Att—across all five datasets and all three evaluation dimensions. This further supports the necessity and synergistic effect of each feature module and the multi-level attention architecture.



(a)



(b)

Figure 10. Cont.

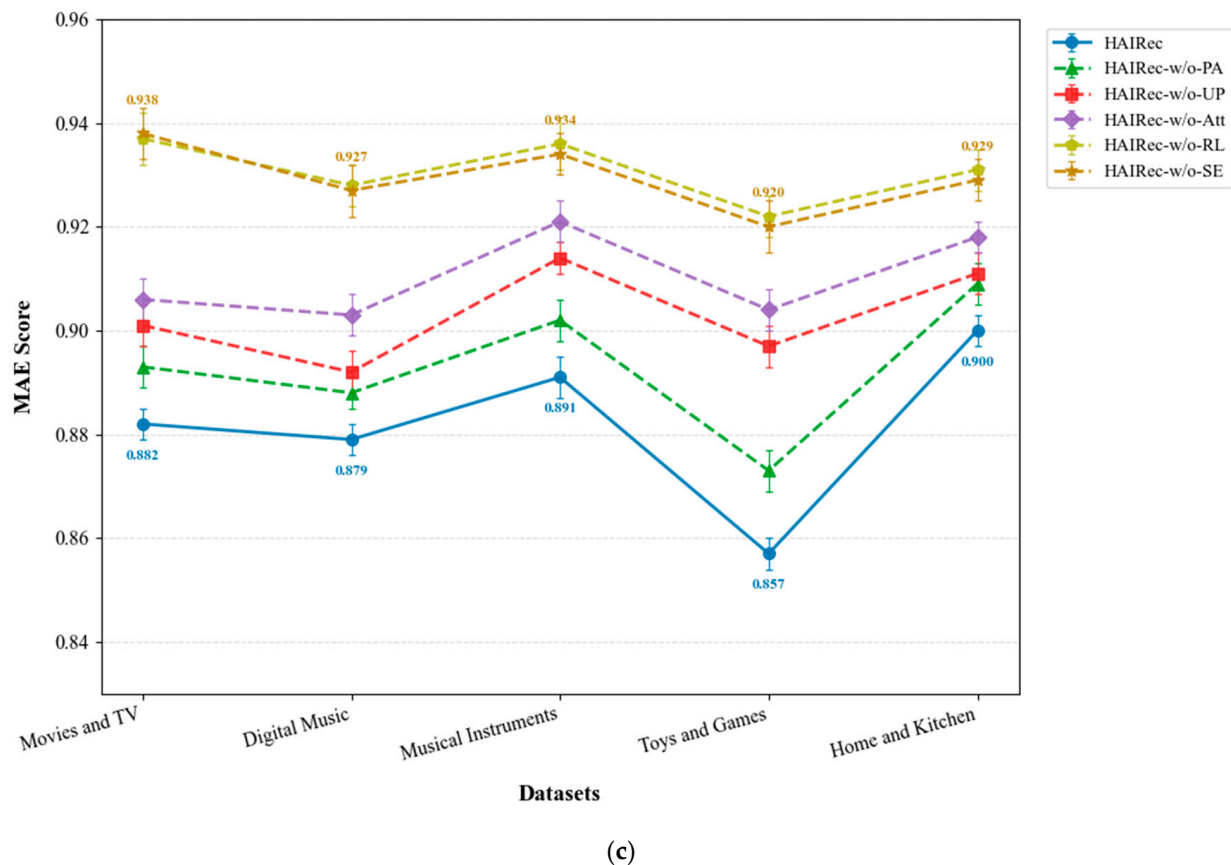


Figure 10. Performance degradation and sensitivity analysis of system-level ablation variants across five real-world datasets: (a) MSE scores; (b) RMSE scores; (c) MAE scores.

Among these variants, HAIRec-w/o-PA, which removes the product attribute extraction module, shows the smallest performance gap from the full model in Figure 10a–c, suggesting that product metadata contributes relatively less to overall accuracy. In contrast, HAIRec-w/o-SE, which omits the sentiment analysis module, exhibits the largest degradation across all three metrics. This demonstrates that sentiment analysis significantly enhances the model’s ability to capture users’ true preferences from rating data. When the rating latent feature (HAIRec-w/o-RL) or user preference module (HAIRec-w/o-UP) is removed, the prediction accuracy drops across all datasets, indicating their essential role in modeling behavioral and interest-level patterns.

Furthermore, replacing the multi-level attention-based feature interaction module with a standard multilayer perceptron (HAIRec-w/o-Att) results in an additional performance decline. These findings highlight that the hierarchical attention structure adaptively reallocates feature importance, mitigates the “information dilution” issue of conventional MLPs, and improves robustness and generalization in high-dimensional sparse environments.

Furthermore, to explicitly quantify the standalone contribution of the error-driven dynamic thresholding strategy within the attention infrastructure, the fine-grained variant HAIRec-w/o-DM is isolated and evaluated via controlled test benchmarks. The empirical tracking indicates that upon bypassing this proactive token suppression gate, a steady performance degradation is observed uniformly across all evaluation environments (for instance, yielding a noticeable average increase of approximately 1.4% to 1.8% in terms of MSE error compared to the full HAIRec model). This targeted empirical decline strictly verifies that the dynamic masking strategy successfully acts as an independent self-supervised regularizer, effectively mitigating cognitive tunneling, eliminating hyper-emotional token noise, and forcing the network to explore more resilient alternative semantic propagation paths.

4.5. Sub-Component Analysis of the LT Module

To further justify the fine-grained design choices within the LT module (consisting of LDA, K-means, and PCA), which is presented as a primary contribution of this work for handling feature heterogeneity and review sparsity, a sub-component ablation study was conducted on the representative Movies & TV dataset. Three inner-module variants were designed for a deeper comparative analysis:

- (1) HAIRec-LT-w/o-K: Omits the K-means clustering layer within the sentiment analysis pipeline, directly feeding the raw, unaggregated topic distributions generated by LDA into the subsequent space-alignment and projection layer.
- (2) HAIRec-LT-w/o-PCA: Replaces the unsupervised PCA dimensionality reduction matrix with a standard linear projection layer (MLP) for cross-space latent alignment within the sentiment module.
- (3) HAIRec-w/o-LT: Completely removes the entire LT-based sentiment analysis module (the data are reused from Section 4.4 as the baseline performance when the LT module is fully absent).

The experimental results across all three evaluation metrics are summarized in Table 6.

Table 6. Sub-component ablation results of the LT module on Movies & TV dataset.

Model Variant	MSE	RMSE	MAE
HAIRec-w/o-LT	0.939 ± 0.007	0.969 ± 0.005	0.937 ± 0.004
HAIRec-LT-w/o-K	0.915 ± 0.005	0.957 ± 0.004	0.912 ± 0.003
HAIRec-LT-w/o-PCA	0.899 ± 0.005	0.948 ± 0.004	0.894 ± 0.003
HAIRec (Proposed)	0.884 ± 0.004 *	0.940 ± 0.003 *	0.882 ± 0.002 *

As observed from Table 6, omitting any sub-component within the LT module leads to a noticeable performance degradation compared to the full HAIRec model, confirming that each algebraic and statistical step is essential. Specifically, when K-means clustering is removed (HAIRec-LT-w/o-K), the prediction error increases significantly. This demonstrates that applying K-means to aggregate latent topics can effectively compress sparse semantic representations and filter out redundant review noise.

Furthermore, replacing PCA with a standard linear layer (HAIRec-LT-w/o-PCA) also yields inferior results, which implies that the unsupervised geometric projections of PCA maintain more robust and generalized low-dimensional structures before cross-space interaction. In summary, the collaboration of LDA, K-means, and PCA provides an optimal pipeline for latent topic extraction and emotional alignment, firmly validating the architecture design of the proposed HAIRec framework for the LT module.

4.6. Computational Complexity Analysis

To evaluate the practical efficiency and scalability of the proposed HAIRec framework, a quantitative computational cost analysis was conducted by comparing trainable parameters, training time per epoch, and inference latency with baseline models. To help conduct a fair comparison, all measurements were performed on the same representative Movies and TV dataset under identical experimental settings. The detailed results are summarized in Table 7.

Table 7. Quantitative computational complexity comparison under identical experimental settings on the Movies and TV dataset.

Model	Trainable Parameters (M)	Training Time/Epoch (s)	Inference Latency (ms/Sample)
PMF	1.82	11.8	0.06
HFT	4.27	39.6	0.42
ConvMF	6.54	71.3	0.78
DeepCoNN	8.91	118.5	1.36
D-Attn	10.18	132.7	1.58
NARRE	9.46	124.6	1.47
LightGCN	3.36	31.4	0.31
HAIRec (Proposed)	3.59	15.4	0.18

As shown in Table 7, traditional matrix factorization methods achieve lower computational costs due to their simple structures, while review-aware deep models introduce additional overhead from complex text representation learning. Although HAIRec incorporates semantic representations from a pretrained language model, the text embeddings are generated offline during preprocessing, avoiding additional encoding costs during training and inference.

HAIRec contains approximately 3.59 M trainable parameters on the Movies and TV dataset, requiring only 15.4 s per training epoch and achieving an inference latency of 0.18 ms per sample. These results demonstrate that HAIRec achieves an effective balance between recommendation accuracy and computational efficiency.

5. Discussion and Conclusions

This study proposes a hybrid recommendation framework (HAIRec) that integrates review sentiment analysis with user–item interaction data to overcome the inherent limitations of conventional recommender systems in textual representation and multi-source interaction modeling. By extracting fine-grained emotional semantics, latent ratings, user preferences, and product attributes, and leveraging a dual-level attention mechanism, the framework achieves dynamic cross-domain feature fusion and multi-granularity preference learning. Combined with DCN-v2 and MLP components, HAIRec effectively captures both explicit and implicit high-order feature interactions, significantly improving recommendation accuracy and generalization capability.

Beyond the empirical improvements, the findings provide deeper insights into the design of review-aware recommender systems. Previous representative methods, such as DeepCoNN and NARRE, demonstrated the effectiveness of incorporating review information into recommendation, yet their feature fusion strategies mainly relied on shallow concatenation or single-level attention. In contrast, the consistently superior performance of HAIRec across multiple benchmark datasets indicates that explicitly separating module-level feature importance from fine-grained semantic dependencies enables more effective heterogeneous representation learning. Furthermore, compared with graph-based recommendation methods such as LightGCN, which primarily exploit structural collaborative signals, HAIRec jointly models behavioral interactions and review semantics through explicit and implicit high-order feature interactions. These results suggest that semantic–behavioral alignment, together with hierarchical feature interaction, plays a crucial role in improving recommendation robustness under sparse data conditions.

The proposed framework also offers broader implications for recommendation research. The hierarchical heterogeneous interaction mechanism provides an effective

paradigm for integrating unstructured textual semantics with structured behavioral information, which can be naturally extended to other multimodal recommendation scenarios involving textual, visual, or audio content. Moreover, the separation of offline semantic representation learning and lightweight online recommendation makes the framework computationally practical, underscoring its applicability to large-scale industrial systems.

Despite these advantages, HAIRec has several limitations that warrant attention. First, the framework heavily depends on the quality and authenticity of user-generated reviews. Although the proposed text encoder and masking attention mechanism effectively capture fine-grained emotional information, their efficacy may diminish when reviews contain sarcasm, implicit metaphors, or contradictory sentiment expressions. Under such circumstances, inconsistencies between textual semantics and rating signals can introduce optimization conflicts during multi-task learning, leading to suboptimal latent representations.

Second, HAIRec experiences performance degradation under extreme long-tail and absolute cold-start scenarios. Although the hierarchical attention mechanism effectively alleviates moderate data sparsity, it still relies on historical review information to construct meaningful semantic representations. For newly registered users or newly launched items lacking textual interactions, both the sub-feature attention and module-level attention modules lack sufficient semantic inputs. Consequently, the feature interaction module gradually degenerates toward relying mainly on rating information, reducing the effectiveness of heterogeneous feature learning.

Overall, empirical evaluations across five real-world Amazon benchmark datasets demonstrate that HAIRec effectively resolves semantic granularity inconsistency and insufficient heterogeneous feature interaction in sparse review-aware recommendation. By unifying hierarchical dual-level attention with explicit–implicit high-order feature interaction learning, the proposed framework consistently enhances recommendation accuracy and robustness. These findings validate HAIRec as a compelling solution for static review-aware rating prediction under sparse interaction conditions and establish a solid foundation for future extensions to more complex recommendation scenarios.

To address these limitations, future research will explore two complementary directions. First, Heterogeneous Graph Neural Networks (HGNNs) will be incorporated to propagate collaborative signals over high-order relational structures, thereby compensating for the absence of local textual information. Second, open-source LLMs will be integrated through parameter-efficient fine-tuning (PEFT) to generate high-quality pseudo-review embeddings as semantic priors for newly introduced users and items. These extensions are expected to further improve recommendation robustness under extreme sparsity and absolute cold-start conditions while preserving the computational efficiency of the proposed framework.

Author Contributions: Conceptualization, T.G. and G.D.; data curation, T.G.; visualization, T.G.; writing—original draft preparation, T.G.; writing—review and editing, T.G. and G.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets analyzed during the current study are publicly available from the Amazon Review Dataset repository maintained by McAuley Lab. The data can be accessed at <https://nijianmo.github.io/amazon/index.html> (accessed on 9 June 2025). No new datasets were generated during this study. Processed data are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This appendix presents the complete adjusted p -values obtained after applying the Holm–Bonferroni correction. The results cover all comparisons between HAIRec and seven baseline models across five datasets and three evaluation metrics, providing additional evidence for the statistical significance of the reported improvements.

Table A1. Adjusted p -values of pairwise comparisons between HAIRec and baseline models after Holm–Bonferroni correction.

Dataset	Metric	PMF	HFT	ConvMF	DeepCoNN	D-Attn	NARRE	LightGCN
Movies and TV	MSE	0.0012	0.0015	0.0021	0.0024	0.0028	0.0043	0.0068
	RMSE	0.0010	0.0014	0.0019	0.0022	0.0027	0.0041	0.0065
	MAE	0.0013	0.0018	0.0025	0.0029	0.0032	0.0048	0.0071
Digital Music	MSE	0.0015	0.0019	0.0027	0.0031	0.0038	0.0052	0.0074
	RMSE	0.0013	0.0017	0.0024	0.0029	0.0035	0.0050	0.0072
	MAE	0.0016	0.0020	0.0028	0.0033	0.0039	0.0054	0.0078
Musical Instruments	MSE	0.0011	0.0016	0.0023	0.0027	0.0031	0.0046	0.0069
	RMSE	0.0010	0.0015	0.0022	0.0026	0.0030	0.0045	0.0067
	MAE	0.0014	0.0019	0.0026	0.0029	0.0036	0.0051	0.0075
Toys and Games	MSE	0.0012	0.0017	0.0025	0.0029	0.0034	0.0049	0.0071
	RMSE	0.0011	0.0016	0.0024	0.0028	0.0032	0.0047	0.0068
	MAE	0.0015	0.0021	0.0027	0.0032	0.0037	0.0053	0.0076
Home and Kitchen	MSE	0.0013	0.0018	0.0026	0.0029	0.0035	0.0051	0.0073
	RMSE	0.0012	0.0017	0.0026	0.0025	0.0034	0.0049	0.0073
	MAE	0.0016	0.0021	0.0029	0.0034	0.0040	0.0055	0.0095

References

1. Sachdeva, T.; Goyal, L.M.; Mittal, M. Mapping the Landscape of Personalization: A Comprehensive Review of Prediction and Trends in Recommendation Systems. *WIREs Data Min. Knowl. Discov.* **2025**, *15*, e70006. [\[CrossRef\]](#)
2. Liu, H.; Ismail, F.R.; Zhang, W.; Zou, P.; Hussain, T.; Sharma, Y.K.; Lilhore, U.K.; Simaiya, S.; Tekeste, L.G. A scalable hybrid framework for boosting customer experience and operational efficiency in e-commerce. *Sci. Rep.* **2026**, *16*, 8042. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Mittameedi, S.K.; Dogra, V.; Sayal, A. Customer Experience in E-Commerce: A Systematic Review of Metrics, Models, and the Role of AI. *IEEE Access* **2025**, *13*, 154422–154437. [\[CrossRef\]](#)
4. Behera, G.; Nain, N.; Soni, R.K. Integrating user-side information into matrix factorization to address data sparsity of collaborative filtering. *Multimed. Syst.* **2024**, *30*, 64. [\[CrossRef\]](#)
5. Yuan, H.L.; Hernandez, A.A. User Cold Start Problem in Recommendation Systems: A Systematic Review. *IEEE Access* **2023**, *11*, 136958–136977. [\[CrossRef\]](#)
6. Li, Y.J.; Tang, J.; Gonul, S.; Barbe, D. A Novel Approach for Capturing Consumer Behavior and Preference—MulVAEK. *IEEE Comput. Intell. Mag.* **2025**, *20*, 33–44. [\[CrossRef\]](#)
7. Dogra, V.; Mittameedi, S.K. Aspect-Based Approaches for Measuring Customer Feedback in the E-Commerce Industry. In Proceedings of the 2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 10–12 July 2024; pp. 479–484.
8. Miah, M.S.U.; Kabir, M.M.; Bin Sarwar, T.; Safran, M.; Alfarhood, S.; Mridha, M.F. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Sci. Rep.* **2024**, *14*, 9603. [\[CrossRef\]](#) [\[PubMed\]](#)
9. He, Y.; Du, Y.H.; Pu, X.F. Personalized Recommendation System of E-Commerce in the Digital Economy Era: Enhancing Social Connections with Graph Attention Networks. *Appl. Artif. Intell.* **2025**, *39*, 2487417. [\[CrossRef\]](#)
10. Kang, W.C.; McAuley, J. Self-Attentive Sequential Recommendation. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 197–206.

11. Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; Jiang, P. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, China, 3–7 November 2019; pp. 1441–1450.
12. Liu, S.; Wei, L.; Lv, S.; Li, M. Stability and generalization of ℓ_p -regularized stochastic learning for GCN. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, Macao, China, 19–25 August 2023; p. 631.
13. Li, Z.-Y.; Chen, M.-S.; Gao, Y.; Wang, C.-D. Signal Contrastive Enhanced Graph Collaborative Filtering for Recommendation. *Data Sci. Eng.* **2023**, *8*, 318–328. [\[CrossRef\]](#)
14. Zhang, L.; Liu, P.; Gulla, J.A. Recommending on graphs: A comprehensive review from a data perspective. *User Model. User-Adapt. Interact.* **2023**, *33*, 803–888. [\[CrossRef\]](#)
15. Jahanbin, K.; Chahooki, M.A.Z. Aspect-level sentiment analysis in social media using hybrid deep transfer learning approach. *Knowl.-Based Syst.* **2025**, *330*, 114125. [\[CrossRef\]](#)
16. Mahmoudi, A.; Jemielniak, D.; Ciechanowski, L. Assessing Accuracy: A Study of Lexicon and Rule-Based Packages in R and Python for Sentiment Analysis. *IEEE Access* **2024**, *12*, 20169–20180. [\[CrossRef\]](#)
17. Hua, Y.C.; Denny, P.; Wicker, J.; Taskova, K. A systematic review of aspect-based sentiment analysis: Domains, methods, and trends. *Artif. Intell. Rev.* **2024**, *57*, 296. [\[CrossRef\]](#)
18. Liu, P.; Zhang, L.; Gulla, J.A. Pre-train, Prompt, and Recommendation: A Comprehensive Survey of Language Modeling Paradigm Adaptations in Recommender Systems. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1553–1571. [\[CrossRef\]](#)
19. Salakhutdinov, R.; Mnih, A. Probabilistic Matrix Factorization. In Proceedings of the 21st International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007; pp. 1257–1264.
20. McAuley, J.; Leskovec, J. Hidden factors and hidden topics: Understanding rating dimensions with review text. In Proceedings of the 7th ACM Conference on Recommender Systems, Hong Kong, China, 12–16 October 2013; pp. 165–172.
21. Zheng, L.; Noroozi, V.; Yu, P.S. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, UK, 6–10 February 2017; pp. 425–434.
22. Kim, D.; Park, C.; Oh, J.; Lee, S.; Yu, H. Convolutional Matrix Factorization for Document Context-Aware Recommendation. In Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, 15–19 September 2016; pp. 233–240.
23. Seo, S.; Huang, J.; Yang, H.; Liu, Y. Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction. In Proceedings of the Eleventh ACM Conference on Recommender Systems, Como, Italy, 27–31 August 2017; pp. 297–305.
24. Chen, C.; Zhang, M.; Liu, Y.; Ma, S. Neural Attentional Rating Regression with Review-level Explanations. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 1583–1592.
25. Chin, J.Y.; Zhao, K.; Joty, S.; Cong, G. ANR: Aspect-based Neural Recommender. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Torino, Italy, 22–26 October 2018; pp. 147–156.
26. Tay, Y.; Luu, A.T.; Hui, S.C. Multi-Pointer Co-Attention Networks for Recommendation. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 2309–2318.
27. Wu, L.; Quan, C.; Li, C.; Wang, Q.; Zheng, B.; Luo, X. A Context-Aware User-Item Representation Learning for Item Recommendation. *ACM Trans. Inf. Syst.* **2019**, *37*, 22. [\[CrossRef\]](#)
28. Liu, D.; Li, J.; Du, B.; Chang, J.; Gao, R. DAML: Dual Attention Mutual Learning between Ratings and Reviews for Item Recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 344–352.
29. Wang, X.; He, X.; Wang, M.; Feng, F.; Chua, T.-S. Neural Graph Collaborative Filtering. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 165–174.
30. He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; Wang, M. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020), Virtual Event, 25–30 July 2020; pp. 639–648. [\[CrossRef\]](#)
31. Bao, K.; Zhang, J.; Zhang, Y.; Wang, W.; Feng, F.; He, X. TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation. In Proceedings of the 17th ACM Conference on Recommender Systems, Singapore, 18–22 September 2023; pp. 1007–1014.
32. Gao, Y.; Sheng, T.; Xiang, Y.; Xiong, Y.; Wang, H.; Zhang, J. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. *arXiv* **2023**, arXiv:2303.14524.
33. Maragheh, R.Y.; Vadla, P.; Gupta, P.; Zhao, K.; Inan, A.; Yao, K.; Xu, J.; Kanumala, P.; Cho, J.H.D.; Kumar, S. ARAG: Agentic Retrieval Augmented Generation for Personalized Recommendation. *arXiv* **2025**, arXiv:2506.21931.
34. Wang, R.; Shivanna, R.; Cheng, D.; Jain, S.; Lin, D.; Hong, L.; Chi, E. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 1785–1797.

35. Shan, Y. Social Network Text Sentiment Analysis Method Based on CNN-BiGRU in Big Data Environment. *Mob. Inf. Syst.* **2023**, *2023*, 8920094. [[CrossRef](#)]
36. Cho, K.; Merriënboer, B.V.; Bahdanau, D.; Bengio, Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), Doha, Qatar, 25 October 2014; pp. 103–111.
37. Tan, K.L.; Lee, C.P.; Lim, K.M. RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis. *Appl. Sci.* **2023**, *13*, 3915. [[CrossRef](#)]
38. Liu, F.; Wang, J.; Yang, J.Y. A Graph Neural Network Recommendation Method Integrating Multi Head Attention Mechanism and Improved Gated Recurrent Unit Algorithm. *IEEE Access* **2023**, *11*, 116879–116891. [[CrossRef](#)]
39. Zimmermann, J.; Champagne, L.E.; Dickens, J.M.; Hazen, B.T. Approaches to improve preprocessing for Latent Dirichlet Allocation topic modeling. *Decis. Support Syst.* **2024**, *185*, 114310. [[CrossRef](#)]
40. Zhou, G.; Zhu, X.; Song, C.; Fan, Y.; Zhu, H.; Ma, X.; Yan, Y.; Jin, J.; Li, H.; Gai, K. Deep Interest Network for Click-Through Rate Prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1059–1068.
41. Sedhain, S.; Menon, A.K.; Sanner, S.; Xie, L. AutoRec: Autoencoders Meet Collaborative Filtering. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 111–112.
42. Liang, D.; Krishnan, R.G.; Hoffman, M.D.; Jebara, T. Variational Autoencoders for Collaborative Filtering. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 689–698.
43. Covington, P.; Adams, J.; Sargin, E. Deep Neural Networks for YouTube Recommendations. In Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, 15–19 September 2016; pp. 191–198.
44. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020.
45. Guo, H.; Tang, R.; Ye, Y.; Li, Z.; He, X. DeepFM: A factorization-machine based neural network for CTR prediction. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 1725–1731.
46. Lian, J.; Zhou, X.; Zhang, F.; Chen, Z.; Xie, X.; Sun, G. xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1754–1763.
47. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011.
48. Ni, J.; Li, J.; McAuley, J. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 188–197. [[CrossRef](#)]
49. Liu, Y. UniDNR: A Novel Interaction Graph and Texts Fusion Method for Review-based Recommendation. In Proceedings of the Twentieth International Conference on Intelligent Computing (ICIC 2024), Tianjin, China, 5–8 August 2024. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.