DMCFRec: A Dual-path Multi-modal Collaborative Filtering Model with Implicit Feedback

Lixue Wang, Ye Tao

Abstract—**Recommender systems, as the core engine of information distribution, have evolved from early collaborative filtering to an intelligent paradigm that fuses multimodal data with deep learning techniques. However, extant approaches generally suffer from a series of problems, such as insufficient modelling of implicit feedback signals, limited exploitation of collaborative relationships, and weak fusion of multimodal information. The paper proposes a novel multimodal recommendation model, DMCFRec, which adopts a parallel modelling mechanism of structural and temporal paths, extracts higher-order synergistic structural information through graph convolutional networks, and models the temporal dependencies of user behaviors in combination with an improved Transformer-Hawkes process. The model introduces an implicit event modelling layer, which fuses implicit feedback signals such as click frequency and browsing duration to enhance the expression of users' potential preferences. It also designs a multimodal adaptive fusion mechanism, which dynamically integrates graph structure, temporal features, and implicit behaviours through multi-head attention. Meanwhile, in order to enhance the recommendation effect in sparse scenarios, the model introduces a lightweight collaborative filtering module as an auxiliary supervision. The experimental findings, derived from two publicly accessible datasets, Beauty and ML-1M, demonstrate that DMCFRec enhances the Recall@10 and NDCG@10 metrics by 4.62% and 7.71%, respectively, when compared with established models such as EEDN, NGCF, LightGCN, and SGL (p ¡ 0.01). The analysis further reveals that DMCFRec exhibits substantial advantages in terms of recommendation accuracy and ranking quality. The ablation experiments further validate the critical contribution of each module to the overall performance. In summary, DMCFRec provides an efficient and scalable solution for personalised recommendation in complex scenarios. It demonstrates good theoretical value and application prospects.**

Index Terms—**Keywords: implicit feedback; graph neural networks; collaborative filtering; multimodal fusion; time series recommendation.**

I. Introduction

INthe contemporary digital economy, the development of the Internet information ecosystem has led to the emergence of recommender systems as a pivotal mechanism for the dissemination of information. These systems have gained significant relevance across diverse domains, capable of generating personalised recommendations for consumers[1]. Recommender systems have been identified as a significant medium for the filtration of information and the provision of personalised services, with a wide range of applications across various domains, including e-commerce, content distribution, and social networks. These systems have been demonstrated to effectively mitigate the challenges posed by information overload, enhancing user engagement through the provision of a personalised experience, and propelling business growth[2]. The prevailing recommendation methods are predicated on explicit feedback, such as user ratings or click data, yet these methods frequently prove ineffective in capturing the evolving interests of users due to their dynamic nature. Consequently, the enhancement of the diversity of recommendations has progressively emerged as a pivotal area of focus within the domain of research concerning recommender systems, with the objective of catering to the escalating demand for diversified material life[3]. Concurrently, challenges such as cold start and data sparsity have persisted in the domain of recommender systems. Despite extensive research, effective solutions remain elusive[4], and frequently encounter limitations[5], resulting in diminished effectiveness in scenarios involving new users and items.

In addressing these challenges, researchers have proposed numerous solutions to enhance the system. Collaborative filtering methods provide a preliminary solution to the cold-start problem by constructing similarity matrices between users and items and mining potential correlation information, while the effective use of implicit feedback signals such as browsing length and click frequency provides new perspectives for modelling users' deeper interests. Moreover, with the continuous emergence of multimodal data, including text, image, and behavioural data, the issue of how to adaptively fuse different modal information to make full use of their complementary advantages has become a major focus in the field of recommender system research in recent years. Concurrently, user behaviour is reflected not only as static points of interest, but also accompanied by clear time-dynamic features. This aspect is frequently disregarded by conventional models; however, enhancing temporal modelling capabilities can facilitate the capture of temporal user interest evolution, thereby facilitating more precise prediction. Consequently, the amalgamation of graph structure modelling and temporal modelling, in conjunction with the implementation of advanced techniques such as the Transformer-Hawkes process[6], has emerged as a pioneering research direction within the domain of recommender systems. This approach facilitates a comprehensive and multi-

Manuscript received May 4, 2025; revised September 6, 2025.

This work was supported by National Natural Science Foundation of China (62272093), Social Science Federation project of Liaoning Province (2025lslybwzzkt-100), and Postgraduate education and teaching reform research project of Liaoning Province (LNYJG2024092).

Lixue Wang is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China (e-mail: 848433132@qq.com).

Ye Tao is an Associate Professor at the School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (corresponding author to provide phone: +86-133-0422-4928; e-mail: taibeijack@163.com).

level characterisation of user behaviours.

This study is conducted in this particular context with the objective of overcoming the limitations of existing methods in implicit feedback, collaborative information exploitation and multimodal fusion. The aim is to enhance the overall performance and robustness of recommender systems in complex scenarios. This is achieved by introducing collaborative filtering modules, designing implicit event modelling layers, constructing multimodal adaptive fusion mechanisms and implementing innovative temporal modelling techniques.

From a theoretical standpoint, the field of recommender systems encompasses a multitude of interdisciplinary domains, including data mining, machine learning, and social network analysis, among others. The findings of the research not only serve to enhance the extant theoretical framework of artificial intelligence, but also furnish novel concepts for the modelling of complex systems. In particular, the research on recommender systems has made significant progress in overcoming the limitations of traditional algorithms, facilitating a more profound and multidimensional analysis of user behaviour, particularly in the domains of dynamic capture of user interest and fusion of multimodal information. From a pragmatic standpoint, recommender systems have been extensively utilised in a plethora of domains, including e-commerce, news, video, and social networking. Enterprises stand to benefit from the employment of accurate recommendation algorithms, which facilitate enhanced comprehension of user preferences, elevated user stickiness, and augmented market conversion rates. Concurrently, these enterprises are able to furnish precise data support for the realms of advertising and product customisation. The practical application of recommender systems has been shown to enhance the competitiveness of enterprises whilst simultaneously providing consumers with a more personalised and superior service experience. In the future, the development of recommender systems will be characterised by an increased focus on real-time functionality, enhanced robustness, and the comprehensive integration of cross-domain information. Advances in big data, the Internet of Things and artificial intelligence technologies have led to significant developments in recommendation systems. These systems are capable of achieving more accurate modeling and intelligent decision-making, thereby promoting efficient information circulation and resource allocation in the digital economy. This, in turn, has the potential to enhance the overall level of social informatization.

In recent years, the role of implicit feedback in recommender systems has been the focus of increased attention. Conventional recommendation methodologies predominantly depend on explicit rating data, such as Matrix Factorization (MF)[7] and Collaborative Filtering (CF)[8]. However, in practice, explicit ratings are frequently sparse and challenging to encompass the entire range of user preferences. Consequently, researchers have adopted the utilisation of modelling data, such as users' clicks, browsing behaviour, dwell time, and purchasing behaviour, as an indirect indicator of users' interests. For instance, the Implicit MF[9] model proposed by Hu et al. builds a trust matrix based on the intensity of users' behaviours; He et al. fused implicit signals with a neural network structure in NCF (Neural Collaborative Filtering)[10]; while behavioural sequence-based models such as DIN[11] and DIEN[12] use attentional mechanisms to enhance the portrayal of behavioural preferences. As demonstrated above, the existing methods have enhanced the performance of the models to a certain extent. However, the majority of these methods do not systematically integrate the joint modelling of multiple implicit signals, such as behavioural frequencies and durations. Furthermore, they lack deep integration with temporal dynamics modelling, which results in an insufficient ability to capture the interests of long-term users.

In the context of collaborative relationship modelling and multimodal fusion, conventional collaborative filtering methodologies such as ItemKNN[13] and SVD++[14] have attained certain outcomes by means of explicit modelling of the interaction similarities between users or items. The development of graph neural networks (GNN) has led to significant advancements in recommendation efficiency, with LightGCN[15] being a notable example of this progress. This improvement is achieved by simplifying graph convolution operations (removing nonlinear activations and feature transformations), focusing on neighbourhood aggregation. Furthermore, SGL (Self-supervised Graph Learning)[16] introduces a self-supervised comparison learning task, which optimises graph representations through data augmentation and node contrast to optimise graph representation. HCCF (Hypergraph-enhanced Cross-view Contrastive Learning)[17] further exploits the hypergraph structure to capture higher-order semantic relationships and combines cross-view contrastive learning to enhance the model's generalisation ability. Conversely, NCL (Node Contrastive Learning)[18] strengthens the differentiation between positive and negative samples through node-level contrast learning, which improves the robustness in sparse scenarios. Moreover, EEDN (Enhanced Encoder-Decoder Network)[19] illustrates the merits of multimodal feature integration in POI recommendation by means of the fusion of local interactions with global graph structure information through an encoder-decoder architecture.

Nevertheless, extant approaches are still manifestly limited in scope. Firstly, the majority of graph models, including LightGCN and NGCF[20], neglect to fully consider the implicit feedback differences in user behaviours. Despite their ability to capture higher-order collaborative signals, this results in biased node representations. Furthermore, the fusion mechanisms of sequences and graph structures, such as SR-GNN[21] and TAGNN[22], predominantly adopt static weight assignments. These assignments are challenging to adapt to the dynamic changes of multi-source heterogeneous data, especially in scenarios where data is noisy or cold-start, resulting in significant performance degradation. Additionally, although comparative learning frameworks such as SGL and NCL alleviate the problem of sparsity through data augmentation, the joint modelling of positive and negative feedback in implicit behaviours remains inadequate. For instance, EEDN does not explicitly differentiate between preference strength and time decay effects in user behaviour, despite introducing global interaction features. These shortcomings have a deleterious effect on the accuracy and interpretability of existing models in complex scenarios. Consequently, there is still scope for enhancement in the fine-grained characterisation of implicit feedback modelling, multimodal dynamic fusion mechanisms, and temporal-structural co-optimisation. There is also a significant requirement for

a novel recommender framework that can integrate user behaviour signals in a consolidated manner, adaptively assign modal weights, and enhance the robustness of sparse data.

The purpose of this paper is to propose a dual-path multimodal recommendation model, termed DMCFRec (Dual-path Multi-modal Collaborative Filtering Framework with Implicit Feedback), and to develop an innovative design around the following two key directions:

(1) In the context of implicit feedback modelling, this paper proposes the Implicit Event Layer (IEL), which, for the first time, incorporates implicit behavioural signals, such as the number of user clicks and browsing duration, into the sequence modelling process, instead of relying on explicit ratings or item access sequences. This development significantly enriches the semantic representation of user behaviour. Concurrently, in order to enhance the model's capacity to simulate the temporal patterns of user behaviours, this paper proposes a temporal fusion predictor (TFP) founded on the time-perceived intensity function (Hawkes process). The probability of non-event occurrence is estimated through Monte Carlo integration, facilitating the model's discernment of the distinction between actual and potential non-occurred behaviours concurrently. This enables a more precise depiction of the trend in the evolution of user interest. The model can be used to achieve a more accurate portrayal of the evolutionary trend of user interests.

(2) In the context of collaborative fusion mechanisms, the present paper proposes a structural fusion system comprising a collaborative filtering (CF) module and a multi-modal adaptive fusion (MAF) module. Firstly, the introduction of a cosine similarity matrix between user-user and item-item relationships enables the construction of a lightweight collaborative filtering module. This module has been shown to enhance the robustness of the model's recommendations in cold-start and sparse data scenarios. Secondly, the paper puts forward a proposal for an adaptive fusion strategy based on a multi-modal attention mechanism. This strategy is capable of dynamically assigning different importance weights to different modal features. This, in turn, enhances the recommendation robustness in cold-start and sparse data scenarios. The strategy is also able to provide graph structural information and time-series representations in the user's behavioural sequences. The employment of distinct importance weights has been demonstrated to enhance the fusion capability and generalisation performance of the model when processing heterogeneous information from multiple sources.

The integration of the aforementioned dual-path modelling and collaborative fusion mechanism has been demonstrated to enhance the accuracy and stability of the recommendation results, while preserving model flexibility and scalability.

II. METHODOLOGY

A. Overall structure of the model

Research in the field of recommender systems is characterised by ongoing developments in dynamic interest modelling and multimodal fusion. The graph neural network-based model NFARec[23] is notable for its extraction of long-term dependencies in user behaviour sequences through a self-attention mechanism, and its encoding of co-occurrence relationships between items in conjunction with graph convolutional networks (GCNs). The model captures the global topology of item interactions using adjacency and relevance matrices, and performs end-to-end optimisation through event prediction loss with ranking loss. Nevertheless, the model exhibits notable deficiencies. Firstly, the model depends exclusively on a single deep learning path, which is not combined with the explicit similarity computation of traditional collaborative filtering. This results in underutilisation of global user-item association signals. Secondly, it lacks fine-grained modelling of implicit feedbacks, such as click frequency and browsing time, which makes it difficult to capture short-term preferences. Thirdly, it has a rigid fusion mechanism for multi-modal features, such as time, context, and graph structure, and is unable to dynamically adapt to changes in the data distribution.

In order to address the aforementioned issues, this paper puts forward the DMCFRec framework, which attains systematic innovation through the implementation of a modular design. In the dual-path collaborative filtering module, the model separates and co-optimises the traditional collaborative filtering and deep learning paths for the first time. The conventional collaborative filtering approach involves the calculation of user-user and item-item cosine similarity matrices, with the introduction of a time decay factor to dynamically update weights, thereby addressing the limitation of static similarity calculation. The deep learning approach enhances the original Transformer/THP encoder by integrating short-term click signals in behavioural sequences with long-term interest evolution through the gated attention mechanism. The outputs of the two types of paths are then fused by an adaptive weight assignment module, which takes into account the global stability of collaborative filtering and the local sensitivity of deep learning. In order to further explore the value of multimodal data, the model designs a multimodal dynamic fusion module that integrates user behavioural embedding, temporal embedding, contextual embedding, and collaborative filtering features, and dynamically adjusts the contribution weights of each modality through the gating network. For instance, the temporal embedding layer encodes time bins as low-dimensional vectors to capture the time decay effect, while the context-aware graph convolutional network injects external information such as item categories and prices to enhance topological relationship modelling. Additionally, the model explicitly parses implicit signals in user behaviours through the implicit feedback enhancement module, quantifies the user's recent activity through the click frequency calculator, and filters noisy signals through the gated time-attention mechanism. This highlights the impact of high-confidence behaviours, thus solving the difficulty of preference capture in scenarios where explicit ratings are scarce.

In order to account for the dynamic evolution of user interests, the model proposes a dynamic interest evolution module. This module combines GRU[24] and LSTM[25] networks in order to capture local mutations in short-term behavioural sequences, such as promotion-induced click surges, and slow changes in long-term interests, such as seasonal preference migrations, respectively. The model dynamically aggregates long- and short-term interest representations through a collaborative attention mechanism. Concurrently, the graph structure enhancement module enhances the static graph convolution layer of the original model, designs a behaviour-

aware graph construction strategy, dynamically adjusts the item adjacency matrix weights according to the real-time behaviour of the user, and introduces a time decay factor to weaken the contribution of historical interactions so that the graph structure can be adaptively updated in response to the evolution of user behaviour. The model diagram is shown in Figure 1.

B. Data Preprocessing Module

In order to comprehensively analyse the structured information between users and items, this module initially discretises the raw rating data. The rating threshold is set, and the raw ratings are transformed into positive and negative binary feedbacks. Thereafter, a sparse rating matrix R is generated. The rating matrix is then used to construct the item co-occurrence matrix C, and the adjacency matrix A is generated for subsequent graph modelling. The preprocessing strategy of this module has been shown to enhance the robustness of interaction signals and to provide fundamental support for subsequent graph structure coding.

(1) Scoring matrix discretisation:

$$R_{ui} = \begin{cases} 1, & r_{ui} > \tau \\ -1, & r_{ui} \leq \tau \\ 0, & \textit{User not interacting with the item} \end{cases} \tag{1}$$

In this study,R_{ui} is used to denote the discrete rating of item i by user u It is important to note that r_{ui} is the original rating value (usually 1-5). In addition, τ is the rating threshold, which is always 3.

(2) Item co-occurrence correlation matrix:

$$C = \frac{R^T R}{\max(R^T R)} \tag{2}$$

In this study, the term $C \in R^{|I| \times |I|}$ is employed to denote the normalised item relevance matrix, whilst $R^T R$ is used to represent the number of times each pair of items is jointly rated by the user. The denominator is utilised for maximum value normalisation.

(3) Adjacency matrix construction:

$$A_{ij} = \begin{cases} 1, & \exists u : R_{ui} \neq 0 \wedge R_{uj} \neq 0 \\ 0, & \textit{if not} \end{cases} \tag{3}$$

In this study, A_{ij} is employed to denote the presence of co-occurring interactions between items i and j. In instances where at least one user interacts with both i and j, this is indicated by assigning a value of 1.

C. Event Sequence Coding Module

The behavioural sequences of a user, as recorded over time, are indicative of the trend in that user's evolving preferences. The module utilises embedding vectors and location coding to vectorise the behavioural sequences, introducing a multi-head self-attention mechanism to capture their long-term dependencies and local interest biases. In comparison with the conventional RNN configuration, the self-attention mechanism exhibits superior parallel efficiency and long-term modelling capability, rendering it particularly well-suited for data characterised by sparse sequences. The items in the sequence are initially mapped by embedding matrices and subsequently superimposed with their corresponding position codes. Subsequently, the query, key, and value matrices are obtained through linear mapping and enter the attention mechanism for weight computation. The multi-head attention's output is then fused by splicing and linear transformation into a final representation of the user's historical behaviour as an encoded vector of the user's personality preferences.

(1) Item sequence embedding and location coding:

$$z_{i_k} = e_{i_k} + p_k, \quad k = 1, \ldots, t \tag{4}$$

In this study, the term e_{i_k} is used to denote the embedding vector of item i_k. The term p_k is used to denote the coding vector of the kth position in the sequence, and z_{i_k} is used to denote the final embedding of the fused position information.

(2) Linear transformation of the attention head:

$$Q = ZW^Q, \quad K = ZW^K, \quad V = ZW^V \tag{5}$$

In this study, the item sequence matrix is denoted by Z. The mapping matrices for queries, keys and values are denoted by W^Q, W^K, W^V, respectively.

(3) Single-headed attention mechanisms:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

In this context, d_k denotes the dimension of the key, and $softmax$ is employed to regulate the attention allocation weights.

(4) Multi-attention fusion:

$$h_t = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^O \tag{7}$$

In this context, h_t denotes a composite representation of the user's historical behaviour. The term W^O signifies a multi-head output projection matrix, and H denotes the number of heads.

D. Collaborative Filtering Module

Conventional collaborative filtering methodologies have demonstrated efficacy in the domains of cold-start and long-tail item recommendations, through the utilisation of user similarity metrics. The integration of collaborative filtering information within the deep modelling framework is achieved through the implementation of a collaborative submodule that facilitates neighbourhood score-weighted prediction, utilising the cosine similarity metric between user interaction sets. Specifically, the process commences with the calculation of the similarity between users. Subsequently, the neighbourhood users' ratings are weighted and aggregated based on the similarity to obtain the predicted score. The cosine score will serve as a crucial supplementary feature, facilitating participation in the subsequent fusion process in conjunction with the sequence representation.

(1) User similarity calculation:

$$s(u, v) = \frac{|I_u \cap I_v|}{\sqrt{|I_u| \cdot |I_v|}} \tag{8}$$

In this context, the term I_u is used to denote the set of interaction items pertaining to user u, $s(u, v)$ is employed to represent the cosine similarity.

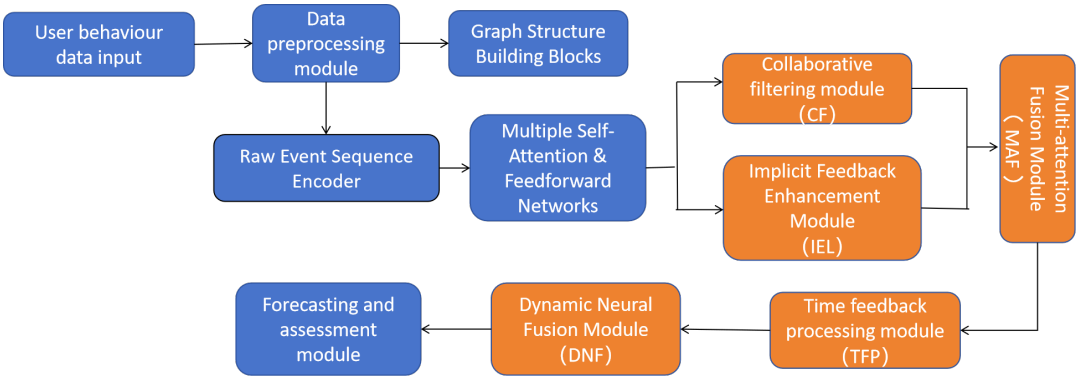

Fig. 1. Diagram of the DMCFRec model

(2) Collaborative filtering score estimation:

$$\hat{r}_{ui}^{\mathrm{CF}} = \sum_{u \in N_u} s(u, v) \cdot R_{vi} \tag{9}$$

The set of nearest neighbours of user u is denoted by N_u, and the rating of item i by nearest neighbour user v is denoted by R_{vi}.

E. Implicit Feedback Enhancement Module

Explicit ratings frequently prove inadequate in providing a comprehensive representation of user preferences. To address this limitation, we propose the integration of implicit behavioural signals, such as the number of clicks and the duration of browsing sessions, as enhancement features. These additions are designed to complement the latent interest exhibited by users during interactions, offering a more holistic perspective on user behaviour. These implicit signals are characterised by their realism and continuity, particularly in circumstances where users demonstrate an absence of active rating behaviour or where ratings are subject to noise. In this module, implicit features are modelled by a multi-layer perceptual machine and fused with user embeddings in a linearly weighted manner to form an augmented user representation that improves the granularity and dynamic adaptability of user interest expression.

(1) Implicit feedback representation:

$$f_{ui}^{\mathrm{imp}} = \phi\left(\mathbf{W}_{\mathrm{imp}} \cdot [\mathbf{c}_{ui}, \mathbf{t}_{ui}] + \mathbf{b}_{\mathrm{imp}}\right) \tag{10}$$

In this equation,$[c_{ui}, t_{ui}]$ is the implicit feedback input vector, W_{imp} and b_{imp} are the MLP (Multi-Layer Perceptron) weights and biases, and $\phi(\cdot)$ is the activation function.

(2) User-embedded enhancements:

$$\tilde{e}_u = e_u + \lambda \cdot f_{ui}^{\mathrm{imp}} \tag{11}$$

In the context of the aforementioned theoretical framework, the term e_u is used to denote the original user embedding, λ is employed to represent the fusion weight, and $\tilde{e}_u$ is used to denote the enhanced user vector.

F. Multi-attention Fusion Module

A plethora of sources can be utilised to ascertain user preference information, including behavioural sequences, collaborative scores and implicit feedbacks, amongst others. The efficacy with which various types of features can be integrated is a critical factor in determining the ultimate performance of the recommender system. The module employs a multi-channel attention mechanism to weight the fusion of multiple feature sources, with each feature channel learning its relative importance through the attention network to avoid information redundancy and conflict. Specifically, the attention score for each feature channel is calculated, and the fusion is weighted according to the attention weights to produce a unified user-item interaction representation vector.

(1) Attention weight calculation:

$$\alpha_j = \frac{\exp(\omega_j^{\top} \tanh(\mathbf{W}\mathbf{x}_j))}{\sum_k \exp(\omega_k^{\top} \tanh(\mathbf{W}\mathbf{x}_k))} \tag{12}$$

In this context, x_j denotes the jth feature channel (e.g., sequence, CF, implicit feedback), and W and w_j represent the attention parameters.

(2) Feature fusion representation:

$$z_{ui} = \sum_j \alpha_j \cdot x_j \tag{13}$$

The final fused interaction representation is denoted by Z_{ui}.

G. Temporal Feedback Modelling Module

Conventional models have historically neglected to consider the impact of temporal information between behaviours on alterations in preference. The Monte Carlo integration approach is utilised in this module to model 'non-event' behaviours, defined as those that do not occur within the specified time span of the user. The module employs this modelling to estimate the effect of the time period on the user's interest, distinguishing between periods of weakening or strengthening. This mechanism enhances the model's capacity to discern short-term interest mutation and long-term interest dilution. The model samples multiple time points within a time interval and models the non-event contributions using the softplus function. This is then fed into the fusion module as a feature complement.

(1) Time-feedback integral estimation:

$$I_{ui}^{\mathrm{time}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{softplus}(h_{ui} + \alpha \cdot \delta t_n) \cdot \Delta t \tag{14}$$

In this study, the term δt_n is employed to denote the nth sampling time point. The parameter α is utilised to denote

the time scaling parameter, whilst h_{ui} is employed to denote the intermediate preference representation. The term Δt is employed to denote the time interval length.

H. Dynamic Fusion and Prediction Module

The function of this module is to integrate the output representations of multiple modules, as previously described, to form a complete feature. The prediction score of the user-item preference is then output through a fully connected neural network with a $Sigmoid$ activation function. The outputs of all modules are spliced into a multidimensional feature vector, which is then nonlinearly transformed and normalised by the neural network to output the prediction score, namely the user u recommended preference for item i.

(1) Feature splicing:

$$\mathbf{z} = \left[\mathbf{h}_t \parallel \hat{r}_{ui}^{CF} \parallel f_{ui}^{imp} \parallel I_{ui}^{time}\right] \tag{15}$$

The symbol || is used to denote the splicing operation, and z is the final feature vector.

(2) Predictive scoring output:

$$\hat{\mathcal{Y}}_{u\bar{i}} = \sigma(\omega^{\top}\mathbf{z} + b) \tag{16}$$

The function $\sigma(\cdot)$ is known as the $Sigmoid$ function, and the predicted probability of the user's preference for the item, $\hat{y}_{ui}$ is defined as belonging to the interval $[0, 1]$.

III. Experimentation

A. Experimental Setting

Two public datasets are utilised in the experimental process: the Amazon beauty category and the MovieLens-1M movie rating dataset. The data are arranged according to timestamp and divided into training, validation and test sets with a ratio of 8:1:1. When ratings ≥ 4 are considered as implicit feedback, the number of user interactions in the Beauty dataset displays a power law distribution (with a significant long-tail effect). In contrast, ml-1M approaches a normal distribution, and the amalgamation of these two approaches provides a comprehensive evaluation of the model's adaptability to varying distribution patterns. In order to eliminate the presence of data bias, the long-tailed distributions of users and items were truncated, and users and items with a minimum of five interactions were retained, as shown in Table I.

TABLE I
STATISTICS OF THE EXPERIMENTAL DATASETS. "#AVG." DENOTES THE AVERAGE COUNT OF USERS' INTERACTIONS. "PERC.(#POS/#NEG)" REFERS TO THE PERCENTAGE OF POSITIVE AND NEGATIVE SAMPLES

Dataset	#Users	#Items	#Interaction	Perc.(#Pos/#Neg)	#Avg
Beauty	22363	12101	198,502	80.2%/19.8%	8.9
MovieLens	6041	3955	1,000,209	73.5%/26.5%	165.6

The model has been implemented using the PyTorch framework, with the optimiser set to Adam, a learning rate of 0.001, a batch size of 2048, and an embedding dimension of 64. In order to prevent overfitting, Dropout (with probability 0.2) and L2 regularisation (with coefficients 1e-4) are utilised. It has been established that, during the training process, the early stopping mechanism is triggered if the validation set loss does not decrease for five consecutive rounds. The model employs the conventional metrics of recall at k and the Normalized Discounted Cumulative Gain (NDCG) at k, which are widely utilised in the domain of recommender systems (k=5, 10).

The recall@k metric is a quantitative assessment of the model's capacity for recall within the context of the Top-k recommendation list, calculated as follows:

$$\text{Recall@k} = \frac{\text{Number of positive sample hits}}{\text{Number of all positive samples of the user}}$$

The NDCG@k metric is employed to evaluate the ranking quality of recommendation lists, with the formula used to calculate this being as follows:

$$\text{NDCG@k} = \frac{\sum_{i=1}^{K} \frac{2^{rel_i}-1}{\log_2(i+1)}}{\text{DCG@k in ideal ordering}}$$

The experiment was replicated on five occasions, and the resulting data were analysed as a mean with standard deviation. Statistical significance was determined by two-tailed Student's t-test, with p-values and confidence intervals (95%) being reported. The calculation of percentage performance improvement was based on mean comparison.

B. Overall Performance Comparison

In order to verify the validity of the models, this paper compares the following classic cutting-edge models:

The EEDN model is a recommendation model that fuses augmented representation learning with deep neural network structure. It employs multi-layer neural networks to model non-linear interactions between users and item embeddings.

LightGCN is a lightweight graph convolutional network with simplified feature transformations and nonlinear activation, emphasising neighbourhood aggregation.

SGL is a self-supervised graph learning framework that enhances data representation through contrast learning.

The HCCF (Hypergraph Collaborative Filtering) model is a data mining technique that exploits the structure of hypergraphs to capture higher-order semantic relationships.

The NCL recommendation model is predicated on the principle of node contrast learning, a methodology that has been demonstrated to enhance the robustness of the model through the augmentation of positive and negative sample comparisons.

As demonstrated by the experimental results presented in the table, as shown in Table II, DMCFRec exhibits a substantial enhancement in recommendation accuracy and ranking quality when compared to the existing baseline model on both datasets. To elaborate, the DMCFRec model's Recall@10 and NDCG@10 values, calculated for the Beauty dataset, demonstrate a 4.62% and 7.71% enhancement over the EEDN baseline, respectively. A similar enhancement is observed for the ml-1M dataset, with Recall@10 and NDCG@10 values of 0.1205 and 0.3341, respectively, indicating an improvement of 6.07% and 16.08% over the EEDN model. It is noteworthy that all these enhancements surpass the statistical significance threshold of $p < 0.01$, as determined by the applied significance test. The experimental

TABLE II
PERFORMANCE COMPARISON BETWEEN DMCFREC AND BASELINE MODEL

Dataset	Metric	EEDN[19]	HCCF[17]	LightGCN[15]	NCL[18]	SGL[16]	SHT[26]	Ours
Beauty	Recall@5	0.0548	0.0367	0.0489	0.0521	0.0522	0.0520	**0.0598**
	Recall@10	0.0779	0.0517	0.0704	0.0688	0.0737	0.0719	**0.0815**
	NDCG@5	0.0476	0.0344	0.0414	0.0453	0.0467	0.0415	**0.0523**
	NDCG@10	0.0558	0.0387	0.0499	0.0502	0.0530	0.0513	**0.0601**
ml-1M	Recall@5	0.0737	0.0603	0.0603	0.0673	0.0665	0.0698	**0.0783**
	Recall@10	0.1136	0.0967	0.0863	0.1029	0.1069	0.1072	**0.1205**
	NDCG@5	0.3222	0.2490	0.2438	0.2579	0.2767	0.2822	**0.3716**
	NDCG@10	0.2878	0.2539	0.2476	0.2497	0.2829	0.2739	**0.3314**

results provide validation of the effectiveness of the dual-path collaborative filtering and multimodal fusion mechanism, especially in sparse scenarios (e.g., Beauty), where the model captures fine-grained behavioural signals through the implicit feedback enhancement module to further alleviate the data sparsity problem.

C. Analysis of ablation experiments

In order to validate the effectiveness of the core modules of the model, the present paper conducts systematic ablation experiments on each of the two datasets, removing the following modules in turn for comparative experiments, as shown in Table III. The Collaborative Filtering Path (w/o CF) denotes the elimination of the direct propagation path for user–item interactions. The Interest Evolution Module (w/o IEL) is characterised by the disablement of temporal feature extraction and the attention mechanism. The Multimodal Fusion Mechanism (w/o MAF) involves the retention of a single modality, such as text or images, instead of integrating multiple data sources. The absence of Temporal Feature Processing (w/o TFP) entails the disregard of temporal context information pertaining to user behaviour.

In the context of the Beauty dataset, the full model attains 0.0815 and 0.0601 for Recall@10 and NDCG@10, respectively. Conversely, the removal of the collaborative filtering path (w/o CF) results in a decline of approximately 7.9% in NDCG@10, indicating that this module plays a pivotal role in enhancing the robustness of recommendations through the mechanism of explicit similarity propagation. Disabling the interest evolution module (without IEL) similarly leads to a performance degradation, with NDCG@10 decreasing from 0.0601 to 0.0543, validating the importance of temporal modelling in capturing users' dynamic preferences. Furthermore, the elimination of the multimodal fusion mechanism (w/o MAF) and temporal feature processing (w/o TFP) has been shown to result in a 9.3% and 9.1% performance degradation, respectively. This suggests that both mechanisms contribute positively to the effective integration of heterogeneous information and temporal context modelling.The experimental results for the Beauty dataset are illustrated in Figure 2.

In the ml-1M dataset, the full model achieves a recall@5 of 0.0783 and a recall@10 of 0.1205. A series of ablation experiments were conducted to ascertain the impact of removing various components on the model's performance. It was found that the removal of the Interest Evolution Module (w/o IEL) had the greatest impact on recall@5, with a significant decrease from 0.0783 to 0.0532, representing a 32.0% reduction. This result further corroborates the importance of the temporal perception mechanism in capturing the evolution of user interests, emphasising the pivotal role of the time-aware mechanism in this process. The elimination of the collaborative filtering path (w/o CF) and the multimodal fusion mechanism (w/o MAF) also results in varying degrees of decrease in the Recall and NDCG metrics, respectively, suggesting that collaborative modelling of multi-source information is of significant importance in enhancing recommendation accuracy. The elimination of the temporal feature processing module (TFP) has also been demonstrated to result in a reduction in the model's capacity to discern temporal context. The experimental outcomes derived from the ml-1M dataset are presented in Figure 3.

In summary, the DMCFRec model has been demonstrated to enhance the performance and stability of the recommender system in complex data environments through the synergistic action of the modules of collaborative filtering paths, interest evolution modelling, multimodal fusion and temporal context processing.

In order to investigate the synergistic effects between different modules, joint ablation experiments were conducted. In these experiments, two core modules were simultaneously removed from the complete model. These modules included collaborative filtering (CF), adaptive multimodal fusion (MAF), interest evolution learning (IEL), and time-aware feature processing (TFP). The experiments were conducted on the *Beauty* and *ML-1M* datasets, with evaluation metrics including Recall@$\{5, 10\}$ and NDCG@$\{5, 10\}$. It is noteworthy that all ablation configurations were maintained at the same level of hyperparameterisation as the complete model. Furthermore, the removed modules were replaced with equidimensional linear mappings in order to circumvent any potential interference that might have arisen from disparities in parameter counts.

The experimental results are displayed in Table IV. In comparison with the complete model, all joint ablations resulted in a decline in performance. Among them, the most significant degradation was observed in the absence of both collaborative filtering (CF) and adaptive multimodal fusion (MAF) (*w/o* CF+MAF). On the *Beauty* dataset, the Recall@10 decreased from 0.0815 to 0.0786, and the NDCG@10 decreased from 0.0601 to 0.0581. On the *ML-1M* dataset, the Recall@10 decreased from 0.1205 to 0.1150, while the NDCG@10 declined from 0.3314 to 0.3129,

TABLE III
ABLATION STUDY ON BEAUTY AND ML-1M DATASETS

Dataset	Beauty				ml-1M	
Metric	Recall@5	Recall@10	NDCG@5	NDCG@10	Recall@5	Recall@10
Full model	0.0598	0.0815	0.0523	0.0601	0.0783	0.1205
w/o CF	0.0547(-8.53%)	0.0753(-7.61%)	0.0478(-8.61%)	0.0553(-7.97%)	0.0748(-4.47%)	0.1150(-4.56%)
w/o IEL	0.0537(-10.2%)	0.0736(-9.69%)	0.0472(-9.75%)	0.0543(-9.65%)	0.0532(-32.06%)	0.0738(-38.76%)
w/o MAF	0.0539(-9.87%)	0.0751(-7.85%)	0.0468(-10.52%)	0.0545(-9.32%)	0.0752(-3.96%)	0.1152(-4.39%)
w/o TFP	0.0534(-10.71%)	0.0752(-7.73%)	0.0469(-10.33%)	0.0547(-8.98%)	0.0745(-4.85%)	0.1153(-4.32%)

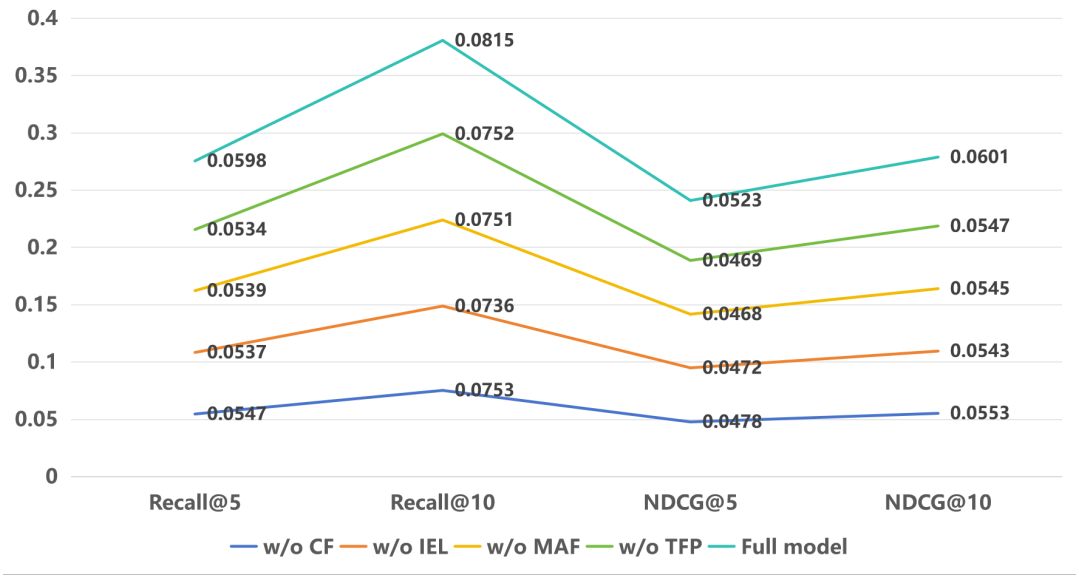

Fig. 2. Comparison of ablation experiments on the Beauty dataset

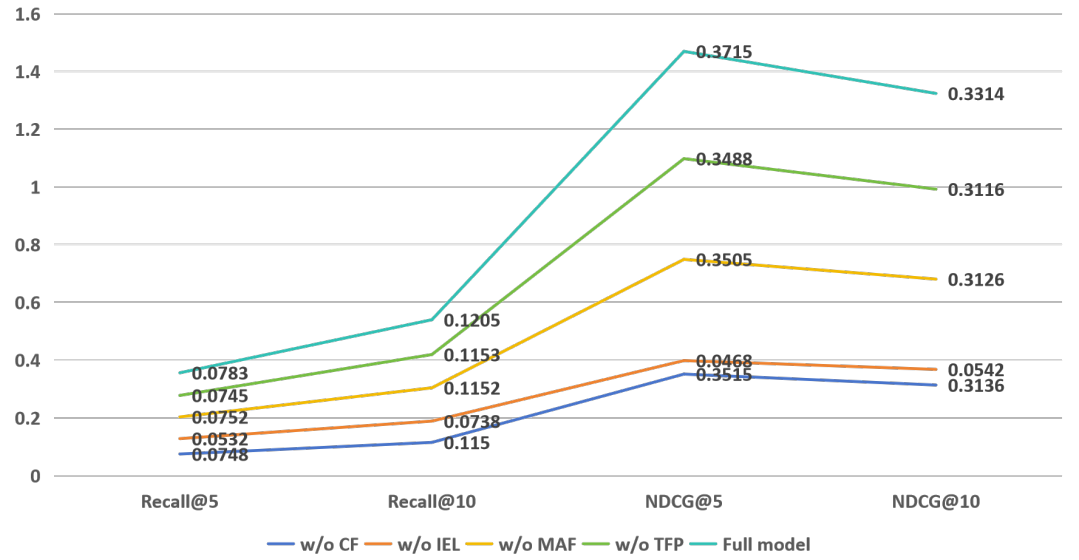

Fig. 3. Comparison of ablation experiments on the ml-1M dataset

TABLE IV
JOINT ABLATION STUDY RESULTS ON BEAUTY AND ML-1M DATASETS.

Metric	*Beauty*				*ML-1M*			
	Recall@5	Recall@10	NDCG@5	NDCG@10	Recall@5	Recall@10	NDCG@5	NDCG@10
W/O CF+MAF	0.0573	0.0786	0.0505	0.0581	0.0758	0.1150	0.3515	0.3129
W/O IEL+TFP	0.0567	0.0776	0.0498	0.0574	0.0755	0.1160	0.3522	0.3142
W/O MAF+TFP	0.0563	0.0774	0.0496	0.0572	0.0745	0.1155	0.3478	0.3111
W/O IEL+MAF	0.0557	0.0775	0.0493	0.0572	0.0761	0.1166	0.3533	0.3150
Full model	**0.0598**	**0.0815**	**0.0523**	**0.0601**	**0.0783**	**0.1205**	**0.3715**	**0.3314**

indicating that the collaborative filtering group-preference prior and the multimodal fusion mechanism have substantial complementary effects in terms of feature alignment and the mitigation of data sparsity. Furthermore, the elimination of MAF and time-aware feature processing (TFP) (*w/o* MAF+TFP), as well as the elimination of interest evolution learning (IEL) and MAF (*w/o* IEL+MAF), also led to noticeable degradation, highlighting that temporal dynamics modelling and interest evolution are indispensable in multimodal information fusion.

D. λlambda Parameter Sensitivity Analysis

This paper conducts a univariate sensitivity analysis to investigate the impact of the λ parameter, which controls the fusion ratio between the enhancement vectors generated from implicit behavioural features such as click frequency, browsing time and the original user embeddings, on model performance. The parameter λ serves as the core hyperparameter in the implicit feedback enhancement module. Different values of λ directly affect the model's ability to balance capturing users' potential interests and suppressing noisy signals.

The experiments involved training and validating six different values of λ ($\{0.0, 0.1, 0.2, 0.5, 1.0, 2.0\}$) on the *Beauty* dataset, while keeping all other hyperparameters consistent with the main experiments. To minimise the influence of randomness, each parameter configuration was evaluated under random seeds $\{0, 1, 2\}$, and the optimal NDCG@10 results were recorded. The final average was then taken as the performance indicator for each configuration.

TABLE V
EXPERIMENTAL RESULTS UNDER DIFFERENT λ VALUES.

λ	Seed=0	Seed=1	Seed=2	Mean ± Standard deviation	
0.0	0.0103	0.0107	0.0105	**0.0105**	±0.0002
0.1	0.0156	0.0163	0.0158	**0.0159**	±0.0004
0.2	0.0206	0.0210	0.0210	**0.0209**	±0.0002
0.5	0.0320	0.0317	0.0323	**0.0320**	±0.0003
1.0	0.0402	0.0405	0.0407	**0.0405**	±0.0003
2.0	0.0473	0.0471	0.0476	**0.0473**	±0.0003

The results in Table V show that as λ increases from 0 to the range of 0.5–1.0, the model's performance steadily improves. This demonstrates that introducing an appropriate amount of implicit feedback features can effectively complement fine-grained information regarding users' interest expression and enhance the model's dynamic adaptability. When $\lambda = 0$, the model degenerates into a baseline structure without implicit feedback enhancement, resulting in significantly lower performance compared to the optimal configuration. In contrast, when λ is excessively large (e.g., $\lambda = 2.0$), the NDCG@10 metric decreases due to over-amplification of the implicit feedback signal, which diminishes the contribution of structured and temporal features. This overemphasis leads to a bias towards short-term behavioural patterns in interest representation. Overall, these results verify the sensitivity of λ to model performance and provide a solid basis for selecting an appropriate parameter value in practical deployments.

IV. SUMMARY

The proposed model, designated as DMCFRec, is a dual-path multimodal collaborative filtering recommendation model intended for complex recommendation scenarios. The aim of the model is to address the core problems of traditional recommendation systems, namely insufficient implicit feedback modelling, limited use of collaborative information, and weak multimodal feature fusion capability. The model introduces a collaborative modelling mechanism between structural and temporal paths, and combines a graph convolutional network and an improved Transformer-Hawkes process to effectively capture higher-order structural relationships and time-dependent features of user behaviours. Concurrently, the design of an implicit feedback enhancement module and a multimodal adaptive fusion mechanism enables the exploration of implicit behavioural signals, such as clicking frequency and browsing time, and contextual features, thereby significantly improving the representation of users' interests. This development has been shown to enhance the granularity and robustness of user interest representation. A series of comparison experiments were conducted on two representative datasets, Beauty and MovieLens-1M, to assess the performance of DMCFRec in relation to existing mainstream recommendation models. The experimental results demonstrated that DMCFRec exhibited superior performance in key metrics such as Recall@10 and NDCG@10. This observation serves to substantiate the effectiveness and validity of DMCFRec under diverse conditions characterised by sparsity and variations in data domains. Subsequent experimentation has demonstrated that the proposed sub-modules fulfil pivotal functions in optimising model performance, with particular emphasis on the collaborative filtering path and implicit feedback modelling modules, which have been shown to enhance accuracy and ranking ability to a significant degree. Consequently, DMCFRec not only enhances the theoretical comprehension of implicit feedback and multimodal fusion, but also offers a pragmatic approach to the development of efficient and robust personalised recommender systems. This approach holds considerable promise for future applications and expansion.

REFERENCES

[1] S. B. Díaz, K. Coussement, and A. De Caigny, "From collaborative filtering to deep learning: Advancing recommender systems with longitudinal data in the financial services industry," *European Journal of Operational Research*, Elsevier, 2025.

[2] H. Xiang, W. Fei, R. Ni, and X. Zhang, "A learning-based anomaly detection framework for secure recommendation," *Information Sciences*, vol. 708, pp. 122071, Elsevier, 2025.

[3] J. Chakraborty and V. Verma, "A survey of diversification techniques in Recommendation Systems," in *Proc. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pp. 35–40, IEEE, 2016.

[4] L. Duan, L. Zhu, and P. Ren, "A Dynamic Cross-Domain Recommendation Model with Target-Aware Complementary Preference Transfer and Information Fusion," *Engineering Applications of Artificial Intelligence*, vol. 148, pp. 110404, Elsevier, 2025.

[5] C. Meng, C. Pan, H. Shu, Q. Wang, H. Guo, and J. Zhu, "Heterogeneous collaborative filtering contrastive learning for social recommendation," *Applied Soft Computing*, vol. 173, pp. 112934, Elsevier, 2025.

[6] S. Zuo, H. Jiang, Z. Li, T. Zhao, and H. Zha, "Transformer hawkes process," in *Proc. International Conference on Machine Learning*, pp. 11692–11702, PMLR, 2020.

[7] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, IEEE, 2009.

[8] D. Yang, Z. T. Nie, and F. Yang, "Time-aware CF and temporal association rule-based personalized hybrid recommender system," *Journal of Organizational and End User Computing*, vol. 33, no. 3, pp. 19–34, IGI Global, 2021.

[9] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proc. 2008 Eighth IEEE International Conference on Data Mining*, pp. 263–272, IEEE, 2008.

[10] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proc. 26th International Conference on World Wide Web*, pp. 173–182, 2017.

[11] G. Zhou et al., "Deep interest network for click-through rate prediction," in *Proc. 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1059–1068, 2018.

[12] G. Zhou et al., "Deep interest evolution network for click-through rate prediction," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 5941–5948, 2019.

[13] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th International Conference on World Wide Web*, pp. 285–295, 2001.

[14] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, IEEE, 2009.

[15] X. He et al., "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648, 2020.

[16] J. Wu et al., "Self-supervised graph learning for recommendation," in *Proc. 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 726–735, 2021.

[17] L. Xia et al., "Hypergraph contrastive collaborative filtering," in *Proc. 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 70–79, 2022.

[18] Z. Lin et al., "Improving graph collaborative filtering with neighborhood-enriched contrastive learning," in *Proc. ACM Web Conference 2022*, pp. 2320–2329, 2022.

[19] X. Wang et al., "EEDN: Enhanced encoder-decoder network with local and global context learning for POI recommendation," in *Proc. 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 383–392, 2023.

[20] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proc. 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 165–174, 2019.

[21] Q. Zhu, N. Ponomareva, J. Han, and B. Perozzi, "Shift-robust GNNs: Overcoming the limitations of localized graph training data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27965–27977, 2021.

[22] F. Yu et al., "TAGNN: Target attentive graph neural networks for session-based recommendation," in *Proc. 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1921–1924, 2020.

[23] X. Wang et al., "NFARec: A negative feedback-aware recommender model," in *Proc. 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 935–945, 2024.

[24] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[25] K. Greff et al., "LSTM: A search space odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, IEEE, 2016.

[26] L. Xia, C. Huang, and C. Zhang, "Self-supervised hypergraph transformer for recommender systems," *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2100–2109, ACM, 2022.