TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214
DOI: 10.26599/TST.2025.9010065

MusReco: A New Transformer-Enhanced Paradigm of Music Recommendation

Duo Xu, Yongsen Zheng, Xin Jin, Leyi Zhao, and Changyin Sun*

Abstract: With the rapid development of the Internet and information technology, the availability of music resources has experienced significant growth. However, along with this abundance, users often find it difficult to navigate through such an extensive collection and locate the specific songs they are looking for, leading to a considerable investment of time and effort. Thankfully, the emergence of music recommendation systems has addressed this problem effectively. These systems leverage advanced algorithms to swiftly help users discover music that aligns with their preferences, thereby saving their valuable time and energy, while also contributing to the economic success of the music platforms. This research centers around the recommendation of music using Transformer-based frameworks. To achieve this, we harness the power of the PyTorch framework to build a comprehensive network model that takes into account essential factors, such as music information, user profiles, contextual details, and historical user behavior data. An efficient transformer module is derived and serves as the backbone of the network model, facilitating the generation of a top-*k* recommendation list tailored to each user's preferences. The module combines the attention free transformer and the convolutional layers. In evaluating our approach, we rely on established metrics, such as accuracy and recall, to assess the level of user interest. Experiments affirm the superiority of our approach, surpassing the performance of existing methods in the field.

Key words: music recommendation; deep learning; Transformer-based model; recommendation system

1 Introduction

The widespread adoption of the Internet has resulted in an unprecedented influx of data and information, leading to a significant challenge known as information overload. To address this issue and alleviate the burden of information overload, focuses have been shifted towards the field of recommendation systems, encompassing various domains such as books, movies, TV programs, and other products[1]. The primary

• Duo Xu is with School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China, and also with Department of Arts Management, Tianjin Conservatory of Music, Tianjin 300171, China. E-mail: xuduotj2024@hotmail.com.
• Yongsen Zheng is with College of Computing and Data Science, Nanyang Technological University, Singapore 639798, Singapore. E-mail: yongsen.zheng@ntu.edu.sg.
• Xin Jin is with Department of Cyberspace Security, Beijing Electronic Science and Technology Institute, Beijing 100070, China. E-mail: jinxin@besti.edu.cn.
• Leyi Zhao is with Luddy School of Informatics, Computing, and Engineering, Indiana University, Indiana Bloomington, IN 47408, USA. E-mail: leyizhao@iu.edu.
• Changyin Sun is with School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China. E-mail: cysun@seu.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2024-11-20; revised: 2025-02-26; accepted: 2025-04-14

objective of a recommendation system is to transform user data and preferences into predictions of their future preferences and interests[2]. Over the years, recommendation systems have experienced rapid growth and development across a diverse range of fields,including the music domain[3]. Today, with an overwhelming abundance of music resources available, users often struggle to efficiently find music that aligns with their interests. In response to this challenge and with the goal of enhancing the user experience, the Music Recommendation System (MRS) has emerged.

Since the inception of Ringo, one of the earliest music recommendation systems developed in 1995[4], the field of music recommendation has undergone rapid development, giving rise to various recommendation methods. These methods have paved the way for advancements in personalized music recommendations and have greatly influenced the growth of the music recommendation domain[5]. As shown in Fig. 1, the typical one is collaborative filtering-based recommendation systems, such as the one proposed by Breese et al.[6], leverage the collective behavior of similar users to make recommendations. More recently, deep learning based[7, 8] recommendation methods, such as the wide and deep model, have gained prominence in music recommendation research. Deep learning models incorporate neural networks that can effectively capture complex relationships and patterns in the music data.

However, the tastes and music preferences of users

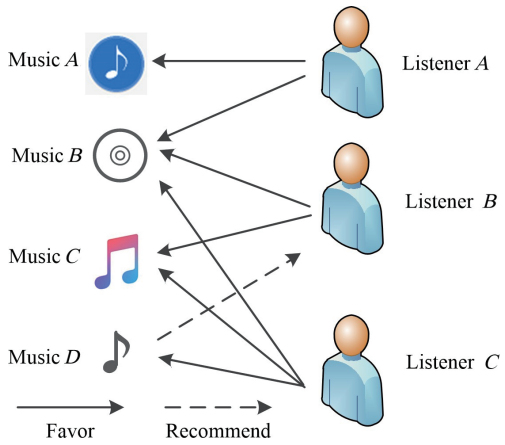


Fig. 1 Collaborative filtering algorithm includes the following three steps: Step 1: Obtain the tag description information of the users; Step 2: Find the users with the common metal model as the target user; Step 3: Produce the initial recommendation dataset from the favorite music of the user with the common mental model.

are influenced by multitude of factors, making it challenging for traditional music recommendation systems to meet users' needs and provide optimal recommendations. Conversely, to effectively cater to users' music entertainment needs, it is crucial to consider several factors, including users' internal characteristics[9], external influences[10], contextual factors[11], and interactive information[12]. For instance, users' internal attributes, such as their personality and emotional state, play a significant role in shaping their music preferences. Additionally, the external factors that influence users, such as their activities and environmental elements like weather conditions, social environment, or historical sites[13], further enhance the accuracy of music recommendations. Furthermore, the generation of music playlists can also take into account specific occasions, providing information on which songs are suitable for the current context[14]. By comprehensively considering all these factors, a music recommendation system can offer more accurate and personalized recommendations, resulting in an enhanced user experience.

To address these issues, we propose a novel transformer-enhanced paradigm of Music Recommendation (MusReco), which combines multiple factors, including input context, music information, user portrait, historical user behavior data, to model diverse user preferences for music recommendation. MusReco focuses on adopting Transformer[15] as the backbone model to encode music emotional characteristics, user mental model, and environmental information, and combine the above information with user historical data to design a music recommendation system. Specifically, MusReco combines all factors into one embedding vector, and adopts convolutional layers to extract correlations among featurs. Such features are fed into a tailored attention free Transformer to learn the in-depth presentations, which are then combined with feed forward networks to output the recommendation results. Evaluation results on multiple benchmarks have verified the superior performance of MusReco compared with several baselines. Generally, the main contribution of this work includes:

- A novel Transformer-based model for music recommendation that covers multiple factors including input context, music information, user portrait, and historical user behavior.
- A novel Transformer structure with attention free

transform to reduce the overhead for efficient representation learning of multi factors in music recommendation system.

- Evaluation on multiple benchmarks to verify the advanced performance of the proposed framework.

2 Related Work

Traditional music recommendation methods include content-based, collaborative filtering based, and hybrid recommendation. In addition, in order to solve the problems such as cold start and data sparsity in MRS, researchers have also adopted methods, such as music recommendation based on deep learning, cross-domain recommendation[16], active learning[17], reinforcement learning[18], conversation based approach[19], context based approach[20], and long tail recommendation[21], to make recommendations more effective. This paper mainly introduces traditional music recommendation methods including music recommendation based on collaborative filtering and music recommendation based on deep learning.

2.1 Music recommendation based on collaborative filtering

The core idea of collaborative filtering is that users' past preference behaviors has a significant impact on their future behaviors, and their previous behavior is basically consistent with their future behavior[22, 23]. Generally speaking, the similarity between users is estimated according to their historical behavior. Then, according to the evaluation of neighbors with high similarity to the target user, predict whether the target user is interested in the project[24]. Such recommendation systems calculate the similarity between users, and predicts projects according to their similar patterns, as is shown in Fig. 1. The user project scoring matrix provides the basis for collaborative filtering technology[25]. Ferraro et al.[26] considered how popularity bias affects collaborative filtering recommendation based on matrix decomposition. At the same time, this paper also indicates that music recommendation algorithms need better evaluation methods, not only limited to user-centered indicators. To deal with the cold-start problem and data sparsity problem, Yoshizaki et al.[27] proposed a music recommendation system that combines collaborative filtering with music recommendation based on impression words. Althbiti et al.[28] proposed a new model, which uses clustering and artificial neural network to solve the problem of data sparsity in collaborative filtering. Besides, Kim et al.[29] standardized user emotional information and collaborative filtering into six categories, and used collaborative filtering to predict user emotional preferences. In the collaborative filtering method proposed by Sanchez-Moreno et al.[30], users' daily listening habits are captured to depict their characteristics and provide more reliable recommendations for users.

2.2 Music recommendation based on deep learning

In recent years, with the development of in-depth learning, new impetus has also been injected into the music recommendation systems[31]. In this field, deep neural networks are used to extract the potential factors of music items from audio signals or metadata, as well as the sequential mode of learning music items (tracks or artists) from music playlists or listening sessions[32]. van den Oord et al.[33] used the user's historical listening data and music's audio signal data to project the user and music into a shared hidden space by combining weighted matrix factorization and convolution neural network, so as to learn the implicit representation of users and songs. Wang et al. [34] proposed a content and context aware music recommendation method based on network embedding, attention mechanisms, and Convolutional Neural Network (CNN). This method can effectively learn music embedding from rich auxiliary information and apply it to recommendation tasks.

Jiang et al.[35] proposed an improved algorithm based on deep neural networks to measure the similarity between different songs. The proposed method enables making suggestions in large systems and comparing the content of songs by "understanding". Zangerle et al.[36] implemented a deep neural network composed of RNN and attention mechanism. The audio feature of user listening history is extracted through scattering transformation, and then the feature and user profile are combined to obtain the recommendation list through the independent cyclic neural network with mixed attention mechanism[37].

Hansen et al.[38] proposed a recursive neural network embedding model (Contextual and Sequential Recurrent Neural Network, CoSeRNN), which can learn users' sequential listening behavior and adapt it to the current environment. The performance of

CoSeRNN in session and tracking recommendation tasks is more than 10% higher than that of the baseline method. Zhang et al.[39] proposed a new Recurrent Conversational Neural Network for Session based Recommendation (RCNN-SR), which makes use of the advantages of GRU and convolutional filter to make session based recommendation. RCNN-SR uses GRU and item level attention mechanism to capture the complex long-term dependency between clicked items, and applies convolution operation process to extract short-term interests and dynamic preferences. Liang et al.[40] proposed a three branch network for music recommendation, one subnet for user preferences, and two subnets sharing parameters for positive and negative terms separately. The goal is that the distance between user preferences and positive items should be closer than that between user preferences and negative items. The positive and negative samples are used to learn the representation and distance measurement between users and items to solve the recommendation task[41].

At the same time, the combination of deep learning algorithm and innovation of Internet of Things (IoTs) technology is applied to the design and construction of intelligent background music system[42]. Quasim et al.[43] believed that the technology of combining emotional maturity with the IoTs system is emerging, and proposed an Emotion-based Music Recommendation and Classification Framework (EMRCF) to classify songs with high accuracy, and track personal interpersonal teams using memory and emotional songs[44]. The proposed innovation prediction accuracy can identify most of the emotional reactions caused by music audience, and effectively classify songs.

3 MusReco Model for Music Recommendation

3.1 Overview

MusReco aims to integrate multi-aspect information including music information, user portrait, context factors, and historical user behavior data for modeling diverse user preferences to benefit music recommendation[45]. The overview of the recommendation system is shown in Fig. 2. In this section, we will introduce the proposed framework, which contains multiple Transformer modules to encode different types of information. Specifically, the proposed framework modifies the multi-head attention in Transformer with Attention Free Transform (AFT) to reduce the model complexity, as is shown in Fig. 3.

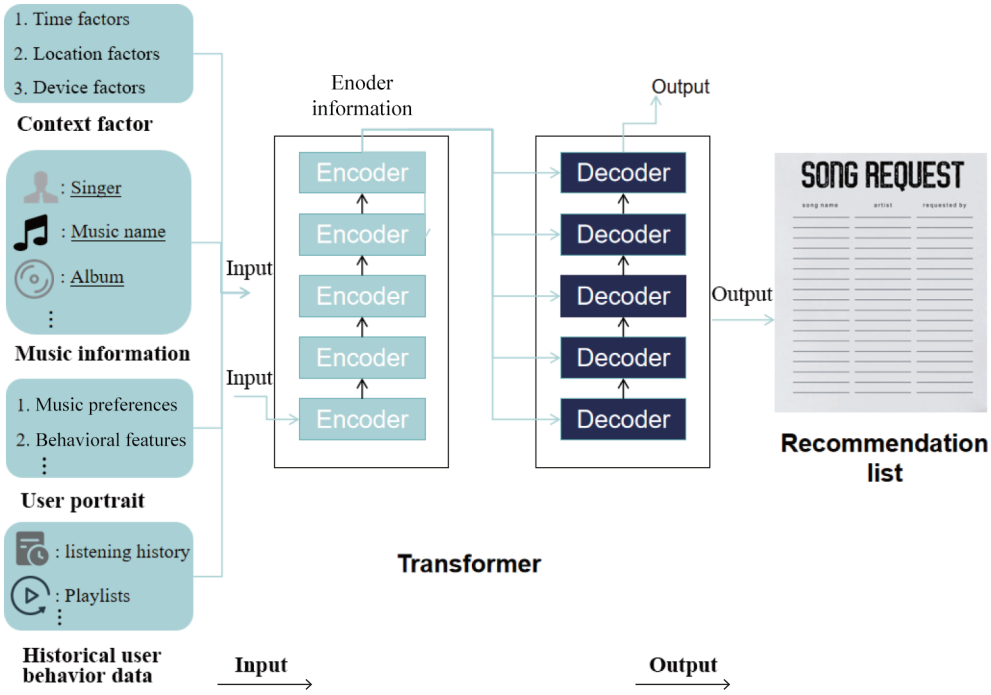


Fig. 2 Overview of MusReco model. The model tasks context factors, music information, user portrait, and historical user behavior data as inputs. These information is encoded and fed into transformer neural network. The output is the recommendation list.

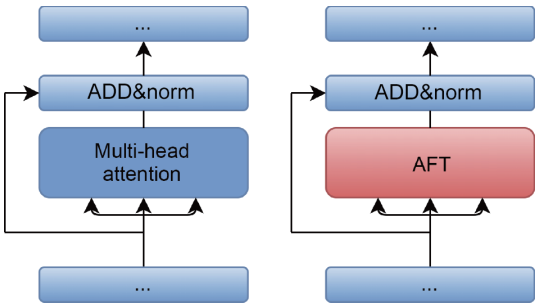


Fig. 3 Original design of Transformer and its modification by replacing multi head attention in Transformer with AFT.

This section first covers the design of Multi-Dconv-Head Attention (MDHA) module, which is the key improvements in the framework. Then the whole model structure for music recommendation is addressed.

3.2 MDHA module

The MDHA module is the core of the recommendation model. It is a Transformer-based layer that can learn representation of multi-aspect information to model diverse user preferences. The main idea of this module is to comprehensively learn feature representations while maintaining the model complexity.

The inputs of this layer include basic music information M, user portrait P, context factors C, and historical user behavior data H. Specifically, music information M includes multi attributes like the genre, artist, and playback time of each piece of music, which can be encoded into vectors. User portrait P includes attributes like gender, personality, and emotional state. The context factors C cover information like the surrounding environment where the user stay. Finally, historical user behavior data H could record the preferred music he/she has listened to.

MDHA first concatenates the feature embeddings of all categories of information as

$$X = [M \oplus P \oplus C \oplus H],$$

and then X is fed into our Transformer. Let $\mathcal{T}^l(X)$ denote the output embeddings from the previous transformer layer, and $\mathcal{T}^{l+1}(X)$ indicates the output of current layer,

$$\mathcal{T}^{l+1}(X) = \mathrm{MDHA}(\mathcal{T}^l(X), \mathcal{T}^l(X), \mathcal{T}^l(X)) \tag{1}$$

where three $\mathcal{T}^l(X)$s refer to key K, query Q, and value matrices V in typical transformers, respectively, as is shown below:

$$\mathrm{MDHA}(K, Q, V) = [h_{l1}; h_{l2}; \ldots; h_{lg}]W_l \tag{2}$$

where g represents the number of heads, W_l denotes the trainable parameters, and each head h_{gl} is calculated by the scaled dot-product attention method represented as SA(·),

$$h_{lg} = \mathrm{SA}(\mathcal{T}^l(X)W_k, \mathcal{T}^l(X)W_q, \mathcal{T}^l(X)W_v) \tag{3}$$

where W_k, W_q, and W_v are assigned as learnable parameters.

To further reduce the complexity of the module, MDHA incorporates the attention free transformer into the module. In the AFT layer, K and V are first combined with a group of learned position deviations, and the results are multiplied by Q in an elemental way. The memory complexity of this operation is linear with the context size and feature dimension, which enables it to adapt to large input and model size at the same time. The specific formula is as follows:

$$Y = \sigma_q(Q) \odot \frac{\sum_{t=1}^{T} \exp(K + \omega_t) \odot V}{\sum_{t=1}^{T} \exp(K + \omega_t)} \tag{4}$$

where Y indicates the output vector of AFT, $\sigma_q(Q)$ is a learnable transformation of Q, "$\odot$" denote the dot product, and ω_t is learnable parameter.

For convenience, we consider the output embeddings of final transformer layer as the final output O,

$$O = \mathrm{MDHA}(\mathcal{T}^L(X), \mathcal{T}^L(X), \mathcal{T}^L(X)) \tag{5}$$

where L is the maximum number of transformer layers.

3.3 Framework for Transformer-enhanced music recommendation

Based on the MDHA module, MusReco model is improved on the basis of the Transformer model. As is shown in Fig. 4, the encoder in the Transformer is divided into three blocks: the first block is input plus location information, the second block is multi head attention, and the third block is feed forward. The second and third blocks have residual connection and layer normalization layers, respectively. The third block is connected to a linear layer and then a softmax layer, which gives the final results for music recommendation. Besides MDHA, MusReco also derives two components for better capturing of features. It adopts the deep convolution layers and the application of the squaring operation to the Rectified Linear Unit (ReLU). The benefits and implications of these modifications are addressed as follow.

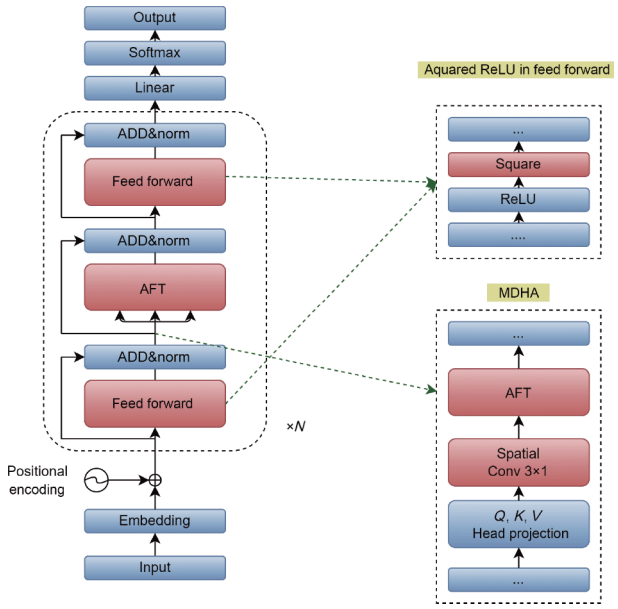


Fig. 4 General diagram illustrating the updated model architecture. This diagram serves as a visual representation of the incorporated changes, showcasing the enhanced components and their relationships within the overall structure (red boxes refer to specific modifications and additions).

First, CNNs have proven to be highly effective in image processing tasks, and their application in recommendation systems has shown promising results as well. By incorporating a deep convolution layer, the model can extract meaningful features and capture intricate patterns from the input data. Moreover, the convolutional layer can explicitly derive correlations across different features. This is especially meaningful for MusReco as it involves multi kinds of information. MusReco adds convolutional layers between the embedding layer and the AFT module, which is a spatial convolution operation with 3×1 convolution kernels,

$$\overline{X} = \text{Conv}_{3\times1}(P(Q,\ K,\ V)) \tag{6}$$

where $P(Q,K,V)$ is the projected embeddings, $\text{Conv}_{3\times1}$ is the convolutional layer, and $\overline{X}$ is the output.

Second, MusReco applies the squaring operation to the ReLU, which transform the activation into a nonlinear function. This alteration can facilitate a better representation of complex relationships between features, leading to enhanced model performance.

To further enhance the model's performance, an additional Feed-Forward Network (FFN) can be stacked, and a full residual link can be placed before the multi-head attention module. This modification introduces a deeper architecture and enables the model to learn more complex representations of the input data. By including the residual link, the model can circumvent the vanishing or exploding gradient problems commonly encountered in deep neural networks. This helps to reduce approximation errors and allows for more accurate predictions.

Finally, we adopt the fully-connected feed-forward network to the final output, which can be modeled as

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \tag{7}$$

In this equation, the learnable weight matrices are denoted by W_1 and W_2, while the bias terms for the first and second layers are represented by b_1 and b_2, respectively.

Efficiency. A notable advantage of incorporating these modifications is the substantial improvement in sample efficiency. Sample efficiency refers to the ability of a model to achieve satisfactory results using fewer training steps. By reducing the number of required training steps, the overall computational burden is reduced, thus making the model more scalable and computationally efficient.

Moreover, the reduction in training steps does not compromise the desired level of quality in the model's outputs. This improvement in efficiency can significantly benefit recommendation systems, as they often operate on large-scale datasets with millions or even billions of user-item interactions. The ability to achieve the desired quality of recommendations with fewer training steps allows for a more rapid deployment of the system and a quicker response to changing user preferences.

4 Experiment

4.1 Benchmarks

A good dataset is crucial to the recommendation algorithm. In this paper, we select four datasets to study the recommendation algorithm, Amazon Digital Music, MSD (Taste Profile subset), Last.fm-1k, and Last.fm-360k[46].

Amazon Digital Music dataset collects 836 006 feedbacks between 478 235 users and 266 414 projects.

MSD is a collection of free audio functions and metadata sets. The core of the dataset is the feature analysis and metadata of one million songs provided by The Echo Nest. When it comes to music datasets, MSD must bear the brunt. The entire dataset is 280 G in size.

Due to the large amount of data, a subset of 10 000 songs is also provided for rapid experiment. In addition, on the basis of MSD dataset, the community also provides seven supplementary datasets for research.

Last.fm-1k and Last.fm-360k are derived from Last.fm, an online radio and music community in UK. It provides developers with rich APIs, which can be called to generate some datasets[47]. Among them, Last.fm-1k dataset contains a representative of the implicit feedback dataset of context information, including music listening records and user information files. The former refers to all music playing records and playing time of relevant listeners, as well as music title, artist name, musicbrain ID, and other information, while the latter records the gender, age, country, and registration time of listeners. Last.fm-360k dataset follows the same organization of Last.fm-1k, which also contains the listening records and user information files. The attributes in each files are identical with those of Last.fm-1k, while the scales of records are extended to 360 000 in total[48].

4.2 Baselines

For the experiment part, the datasets we test are still the four datasets mentioned above (in methodology). At the same time, We select 7 different baseline models to test the effect of our model.

- BPR is a top-k recommendation approach that operates on implicit feedback. It aims to optimize the ranking of items based on user preferences expressed through implicit signals;
- CDAE[49] utilizes a sampling-based learning strategy to uncover the latent representation of corrupted useritem preferences via a denoising autoencoder. It aims to reconstruct the original preferences by training on noisy samples;
- IRGAN[50] is a method that leverages Generative Adversarial Networks (GANs) for recommendation. It consists of a generator, which learns the correlation distribution on the project using signals from the discriminator, and a discriminator uses data selected by the generator;
- CFGAN[51] is a top-k recommendation method based on GANs and sampling learning strategies. It employs vector confrontational learning to provide high-quality recommendations, enhancing the overall recommendation accuracy;
- ENMF[52] is a neural-based matrix decomposition model for top-k recommendation. It effectively learns parameters from the complete training data without sampling, enabling more comprehensive learning;
- TBJE[53] proposes a joint coding method based on Transformers, a popular architecture in natural language processing. The model employs Transformers to capture user-item interactions and generate meaningful representations;
- FairGAN[54] incorporates a fairness-aware learning strategy to dynamically generate fairness signals. By optimizing the search direction, FairGAN explores the space of the best ranking, aiming to distribute item exposure fairly while preserving user utility as much as possible.

4.3 Evaluation metrics

Accuracy of top-k recommendation is a commonly used metric, which is used to measure how many items in the predicted recommendation list are of interest to users. Precision@k calculates how many items are of interest to users in the predicted recommendation list,

$$\text{Precision} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|} \tag{8}$$

where $R(u)$ refers to the given recommendation list, and $T(u)$ refers to the relevant list that users have marked as their favorite.

Top-k recommended recall rate is used to predict the proportion of correct relevant results in all relevant results. Recall@k refers to how many items in the user's real favorite list are predicted by the recommendation algorithm, that is, the recall rate of the real list,

$$\text{Recall} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|} \tag{9}$$

4.4 Experiment results

Based on the experimental results provided in Tables 1–4, we conduct the following analysis and summary for each experiment:

Firstly, in the experiment conducted on the Amazon Digital Music dataset, the performances of several recommendation algorithms are compared. The results show that MusReco achieves the highest results in the Precision@5, Precision@10, Recall@5, and

Table 1 Experimental results on Amazon Digital Music dataset.

(%)

Model	Precision@5	Precision@10	Recall@5	Recall@10
BPR	11.563	15.015	8.964	22.506
CFGAN	12.509	16.513	9.516	24.265
ENMF	12.431	16.399	9.495	24.022
IRGAN	10.428	13.508	8.222	20.688
CDAE	12.120	15.824	9.171	23.184
Fair-GAN	13.326	17.174	10.040	25.208
TBJE	14.633	18.675	11.802	26.532
MusReco	**15.856**	**19.632**	**12.927**	**27.689**

Table 2 Experimental results on MSD (Taste Profile Subset).

(%)

Model	Precision@5	Precision@10	Recall@5	Recall@10
BPR	7.923	10.568	5.564	17.256
CFGAN	8.905	12.872	6.112	19.935
ENMF	8.746	12.756	6.024	19.904
IRGAN	7.359	9.651	4.833	16.233
CDAE	8.526	11.457	5.724	18.821
Fair-GAN	9.654	13.426	6.427	20.368
TBJE	10.826	14.462	7.784	21.656
MusReco	**12.313**	**15.636**	**9.221**	**23.184**

Table 3 Experimental results on Last.fm-1k.

(%)

Model	Precision@5	Precision@10	Recall@5	Recall@10
BPR	15.452	19.762	12.368	25.326
CFGAN	16.589	20.302	13.459	27.584
ENMF	16.325	19.957	13.032	26.328
IRGAN	13.959	17.665	11.753	23.357
CDAE	15.862	19.126	12.946	25.652
Fair-GAN	16.023	21.334	13.757	28.209
TBJE	17.351	22.105	14.816	29.043
MusReco	**18.048**	**23.857**	**15.754**	**29.921**

Recall@10 metrics, with values of 15.856%, 19.632%, 12.927%, and 27.689%, respectively. In the table, bold text indicates the optimal results, while underlined text denotes the suboptimal results. This indicates that MusReco provides more accurate recommendations on this dataset.

Secondly, the experimental results on the MSD (Taste Profile Subset) dataset also demonstrate the outstanding performance of MusReco. MusReco achieves values of 12.313% in Precision@5, 15.636% in Precision@10, 9.221% in Recall@5, and 23.184% in Recall@10. Compared to other algorithms, MusReco significantly improves the accuracy of recommendations and provides more relevant music recommendations for users.

Table 4 Experimental results on Last.fm-360k.

(%)

Model	Precision@5	Precision@10	Recall@5	Recall@10
BPR	5.455	8.359	3.623	15.204
CFGAN	6.386	9.687	4.362	17.631
ENMF	6.124	9.426	4.264	17.459
IRGAN	4.089	7.015	2.524	14.325
CDAE	5.962	8.936	3.876	17.054
Fair-GAN	6.453	10.231	4.828	17.827
TBJE	7.753	11.864	6.636	19.656
MusReco	**8.868**	**12.927**	**7.869**	**20.863**

Thirdly, the experiments conducted on the Last.fm-1k dataset reveal the excellent performance of MusReco. MusReco achieves values of 18.048% in Precision@5, 23.857% in Precision@10, 15.754% in Recall@5, and 29.921% in Recall@10. This again confirms the superiority of MusReco in music recommendation and its higher accuracy in providing personalized recommendations.

Finally, the results of the experiment on the Last.fm-360k dataset further confirm the good performance of MusReco. MusReco achieves values of 8.868% in Precision@5, 12.927% in Precision@10, 7.869% in Recall@5, and 20.863% in Recall@10. This indicates that MusReco can provide accurate music recommendations even on large-scale datasets.

In conclusion, through the analysis of the experimental results on these four datasets, we can conclude that MusReco exhibits the best recommendation performance across various datasets. The results indicate an improvement in precision and recall of 1% to 7% when compared to the other models, which are similar among different datasets. It indicates that MusReco is stable and robust considering the diverse background of these datasets.

Specifically, when compared to Fair-GAN, there is a 2% to 3% increase in both Precision@5 and Precision@10 on all four datasets. Additionally, there is a similar improvement of 2% to 3% in Recall@5 and Recall@10. Similarly, when compared to ENMF, there is a 2% to 3.5% enhancement in Precision@5 and Precision@10, and a 3.5% to 4% increase in Recall@5 and Recall@10 on all four datasets. It outperforms other algorithms in terms of accuracy and personalized

recommendations, providing users with more relevant and satisfying music recommendation experiences. This demonstrates the potential of MusReco in the field of recommendation systems and its ability to offer substantial improvements and advancements in practical applications.

5 Conclusion

This paper proposes a novel music recommendation model, MusReco, which is built upon the transformer-based framework. The model incorporates multi aspects of information correlated with users, music, and contexts to fully cover the factors impacting user preferences. Moreover, the proposed model is improved with an attention free transformer tied with convolutional layers to further improve the learning efficiency and the capabilities for feature representation. According to the experimental results, the performance of MusReco model is better than several state-of-the-art soltuions with similar idea. In the field of music recommendation in the future, we hope to innovate more methods and technologies, and integrate more elements into the recommendation to make music recommendation more rich and diverse, so as to improve user satisfaction and experience.

Acknowledgment

This work was supported by the Ministry of Education Humanities and Social Sciences Research Project on the Emotional Cognition of Synesthetic Effects in Contemporary Chinese Music (No. 20YJC760115).

References

[1] J. Wang, M. J. T. Reinders, J. Pouwelse, and R. L. Lagendijk, Wi-fi walkman: A wireless handheld that shares and recommends music on peer-to-peer networks, in *Proc. SPIE 5683*, *Embedded Processors for Multimedia and Communications II*, San Jose, CA, United States, 2005, pp. 155–163.

[2] P. Resnick and H. R. Varian, Recommender systems, *Commun. ACM*, vol. 40, no. 3, pp. 56–58, 1997.

[3] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas, Music recommender systems, in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, eds, 2nd ed. New York, NY, USA: Springer, 2015, pp. 453–492.

[4] U. Shardanand and P. Maes, Social information filtering: Algorithms for automating “word of mouth”, in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, Denver, CO, USA, 1995, pp. 210–217.

[5] H. T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al., Wide & deep learning for recommender systems, in *Proc. 1st Workshop on Deep Learning for Recommender Systems*, Boston, MA, USA, 2016, pp. 7–10.

[6] J. S. Breese, D. Heckerman, and C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, in *Proc. 14th Conf. Uncertainty in Artificial Intelligence*, Madison, WI, USA, pp. 43–52, 1998.

[7] N. Qiao, L. Dong, and C. Sun, Adaptive deep learning network with multi-scale and multi-dimensional features for underwater image enhancement, *IEEE Trans. Broadcast.*, vol. 69, no. 2, pp. 482–494, 2023.

[8] Q. Fu, K. Lu, and C. Sun, Deep learning aided state estimation for guarded semi-Markov switching systems with soft constraints, *IEEE Trans. Signal Process.*, vol. 71, pp. 3100–3116, 2023.

[9] H. -T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. B. Aradhye, G. Anderson, G. S. Corrado, W. Chai, M. Ispir, et al., Wide & deep learning for recommender systems, in *Proc. 1st Workshop on Deep Learning for Recommender Systems*, https://doi.org/10.1145/2988450.2988454, 2016.

[10] M. S. Mosk, An analysis of music listening behavior as it relates to addiction, undergraduate thesis, Department of Psychology, Tufts University, Medford, MA, USA, http://hdl.handle.net/10427/002140, 2014.

[11] G. Adomavicius and A. Tuzhilin, Context-aware recommender systems, in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, eds. New York, NY, USA: Springer, 2011, pp. 217–253.

[12] S. Kulkarni and S. F. Rodd, Context aware recommendation systems: A review of the state of the art techniques, *Comput. Sci. Rev.*, vol. 37, p. 100255, 2020.

[13] M. Kaminskas, F. Ricci, and M. Schedl, Location-aware music recommendation using auto-tagging and hybrid matching, in *Proc. 7th ACM Conf. Recommender Systems*, Hong Kong, China, 2013, pp. 17–24.

[14] E. Zheleva, J. Guiver, E. Mendes Rodrigues, and N. Milić-Frayling, Statistical models of music-listening sessions in social media, in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, 2010, pp. 1019–1028.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[16] P. Cremonesi, A. Tripodi, and R. Turrin, Cross-domain recommender systems, in *Proc. 2011 IEEE 11th Int. Conf. Data Mining Workshops*, Vancouver, Canada, 2011, pp. 496–503.

[17] D. Carraro, Active learning in recommender systems: An unbiased and beyond-accuracy perspective, PhD dissertation, University College Cork, Cork, Ireland, 2020.

[18] S. Ahmad, J. Zhang, A. Nauman, A. Khan, K. Abbas, and B. Hayat, Deep-EERA: DRL-based energy-efficient resource allocation in UAV-empowered beyond 5G networks, *Tsinghua Science and Technology*, vol. 37, no. 1, pp. 418–432, 2025.

[19] Q. Sun, L. Shi, L. Liu, Z. Han, L. Jiang, Y. Wu, and Y. Zhao, A novel recommendation algorithm integrates resource allocation and resource transfer in weighted bipartite network, *Big Data Mining and Analytics*, vol. 7, no. 2, pp. 357–370, 2024.

[20] K. Haruna, M. Akmar Ismail, S. Suhendroyono, D. Damiasih, A. C. Pierewan, H. Chiroma, and T. Herawan, Context-aware recommender system: A review of recent developmental process and future research direction, *Appl. Sci.*, vol. 7, no. 2, p. 1211, 2017.

[21] H. Yin, B. Cui, J. Li, J. Yao, and C. Chen, Challenging the long tail recommendation, *Proc. VLDB Endow.*, vol. 5, no. 9, pp. 896–907, 2012.

[22] A. Töscher, M. Jahrer, and R. Legenstein, Improved neighborhood-based algorithms for large-scale recommender systems, in *Proc. 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, Las Vegas, NV, USA, 2008, p. 4.

[23] Y. Wang, G. Yin, Z. Cai, Y. Dong, and H. Dong, A trust-based probabilistic recommendation model for social networks, *J. Netw. Comput. Appl.*, vol. 55, pp. 59–67, 2015.

[24] G. Xu, Z. Tang, C. Ma, Y. Liu, and M. Daneshmand, A collaborative filtering recommendation algorithm based on user confidence and time context, *J. Electr. Comput. Eng.*, vol. 2019, no. 1, p. 7070487, 2019.

[25] Y. Shi, M. Larson, and A. Hanjalic, Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges, *ACM Comput. Surveys*, vol. 47, no. 1, p. 3, 2014.

[26] A. Ferraro, D. Bogdanov, X. Serra, and J. Yoon, Artist and style exposure bias in collaborative filtering based music recommendations, arXiv preprint arXiv: 1911.04827, 2019.

[27] S. Yoshizaki, Y. Yoshitomi, C. Koro, and T. Asada, Music recommendation hybrid system for improving recognition ability using collaborative filtering and impression words, *Artif. Life Robot*, vol. 18, no. 1, pp. 109–116, 2013.

[28] A. Althbiti, R. Alshamrani, T. Alghamdi, S. Lee, and X. Ma, Addressing data sparsity in collaborative filtering based recommender systems using clustering and artificial neural network, in *Proc. 2021 IEEE 11th Annu. Computing and Communication Workshop and Conf.* (*CCWC*), Las Vegas, NV, USA, 2021, pp. 218–227.

[29] T. Y. Kim, H. Ko, S. H. Kim, and H. D. Kim, Modeling of recommendation system based on emotional information and collaborative filtering, *Sensors*, vol. 21, no. 6, p. 1997, 2021.

[30] D. Sánchez-Moreno, Y. Zheng, and M. N. Moreno-García, Time-aware music recommender systems: Modeling the evolution of implicit user preferences and user listening habits in a collaborative filtering approach, *Appl. Sci.*, vol. 10, no. 15, p. 5324, 2020.

[31] S. Hong, M. H. Lee, B. S. Yoo, T. Y. Kwak, and S. R. Kim, Application of deep learning algorithms for predicting consolidation settlement, *KSCE J. Civ. Eng.*, vol. 29, no. 1, p. 100072, 2024.

[32] M. Schedl, Deep learning in music recommendation systems, *Front. Appl. Math. Stat.*, vol. 5, p. 44, 2019.

[33] A. van den Oord, S. Dieleman, and B. Schrauwen, Deep content-based music recommendation, in *Proc. 27th Int. Conf. Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2013, pp. 2643–2651.

[34] D. Wang, X. Zhang, D. Yu, G. Xu, and S. Deng, CAME: Content- and context-aware music embedding for recommendation, *IEEE Trans. Neural Network. Learn. Syst.*, vol. 32, no. 3, pp. 1375–1388, 2021.

[35] M. Jiang, Z. Yang, and C. Zhao, What to play next? A RNN-based music recommendation system, in *Proc. 2017 51st Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2017, pp. 356–358.

[36] E. Zangerle, C. M. Chen, M. F. Tsai, and Y. H. Yang, Leveraging affective hashtags for ranking music recommendations, *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 78–91, 2021.

[37] L. Zhang, Y. Wang, K. Yan, Y. Su, N. Alharbe, and S. Feng, Behaviour recognition based on the integration of multigranular motion features in the Internet of Things, *Digit. Commun. Network.*, vol. 10, no. 3, pp. 666–675, 2024.

[38] C. Hansen, C. Hansen, L. Maystre, R. Mehrotra, B. Brost, F. Tomasi, and M. Lalmas, Contextual and sequential user embeddings for large-scale music recommendation, in *Proc. 14th ACM Conf. Recommender Systems*, Virtual Event, 2020, pp. 53–62.

[39] J. Zhang, C. Ma, X. Mu, P. Zhao, C. Zhong, and A. Ruhan, Recurrent convolutional neural network for session-based recommendation, *Neurocomputing*, vol. 437, pp. 157–167, 2021.

[40] H. Liang, D. Zeng, Y. Yu, and K. Oyama, Personalized music recommendation with triplet network, arXiv preprint arXiv: 1908.03738, 2019.

[41] C. Mu, L. Zhang, Z. Wang, Q. Yuan, and C. Peng, Inductive reasoning with type-constrained encoding for emerging entities, *Neural Network.*, vol. 178, p. 106468, 2024.

[42] X. Wen, Using deep learning approach and IoT architecture to build the intelligent music recommendation system, *Soft Comput.*, vol. 25, no. 4, pp. 3087–3096, 2021.

[43] M. T. Quasim, E. H. Alkhammash, M. A. Khan, and M. Hadjouni, RETRACTED ARTICLE: Emotion-based music recommendation and classification using machine learning with IoT framework, *Soft Comput.*, vol. 25, no. 18, pp. 12249–12260, 2021.

[44] Y. Ren, Y. Wang, W. Han, Y. Huang, X. Hou, C. Zhang, D. Bu, X. Gao, and S. Sun, DMSS: An attention-based deep learning model for high-quality mass spectrometry prediction, *Big Data Mining and Analytics*, vol. 7, no. 3, pp. 577–589, 2024.

[45] H. Yin, Y. Liu, Z. Guo, and Y. Wang, From traces to packets: Realistic deep learning based multi-tab website fingerprinting attacks, *Tsinghua Science and Technology*, vol. 30, no. 3, pp. 830–850, 2025.

[46] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, in *Proc. 25th Conf. Uncertainty in*

Artificial Intelligence, Montreal, Canada, 2009, pp. 452–461.

[47] J. Han, W. Li, Z. Cai, and Y. Li, Multi-aggregator time-warping heterogeneous graph neural network for personalized micro-video recommendation, in *Proc. 31st ACM Int. Conf. Information & Knowledge Management*, Atlanta, GA, USA, 2022, pp. 676–685.

[48] X. Min, W. Li, R. Han, T. Ji, and W. Xie, Graph neural collaborative filtering with medical content-aware pre-training for treatment pattern recommendation, *Pattern Recognit. Lett.*, vol. 185, pp. 210–217, 2024.

[49] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester, Collaborative denoising auto-encoders for top-n recommender systems, in *Proc. 9th ACM Int. Conf. Web Search and Data Mining*, San Francisco, CA, USA, 2016, pp. 153–162.

[50] J. Wang, L. Yu, W. Zhang, Y. Gong, Y. Xu, B. Wang, P. Zhang, and D. Zhang, IRGAN: A minimax game for unifying generative and discriminative information retrieval models, in *Proc. 40th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, 2017, pp. 515–524.

[51] D. K. Chae, J. S. Kang, S. W. Kim, and J. T. Lee, CFGAN: A generic collaborative filtering framework based on generative adversarial networks, in *Proc. 27th ACM Int. Conf. Information and Knowledge Management*, Torino, Italy, 2018, pp. 137–146.

[52] C. Chen, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, Efficient neural matrix factorization without sampling for recommendation, *ACM Trans. Inf. Syst.*, vol. 38, no. 2, p. 14, 2020.

[53] J. B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, A transformer-based joint-encoding for emotion recognition and sentiment analysis, in *Proc. 2nd Grand-Challenge and Workshop on Multimodal Language* (*Challenge-HML*), Seattle, WA, USA, 2020, pp. 1–7.

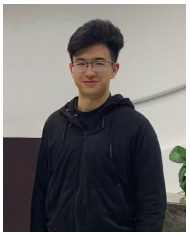
[54] J. Li, Y. Ren, and K. Deng, FairGAN: GANs-based fairness-aware learning for recommendations with implicit feedback, in *Proc. ACM Web Conf. 2022*, Virtual Event, 2022, pp. 297–307.

Duo Xu received the BEng degree in computer science and technology from University of Electronic Science and Technology of China in 2003, and the MEng degree from Fudan University, China in 2009. She is currently an associate professor at Tianjin Conservatory of Music and a research fellow at Beijing General Artificial Intelligence Research Institute, China. Her current research interests include artificial intelligence arts, multimedia technology, and AI-driven art management. She is a senior member of the China Computer Federation and a professional member of the Chinese Musicians Association.

Xin Jin received the BEng degree in computer science from Beijing University of Chemical Technology, Beijing, China in 2006, and the PhD degree in computer science from Beihang University, Beijing, China in 2013. He is currently a professor at Department of Cyber Security, Beijing Electronic Science and Technology Institute, Beijing, China. His research interests include computational aesthetics and AI art.

Changyin Sun received the BS degree from Sichuan University, Chengdu, China in 1996, and the MEng and PhD degrees in electrical engineering from Southeast University, Nanjing, China in 2001 and 2003, respectively. He is currently a professor at School of Artificial Intelligence, Anhui University, Hefei, China, and also a professor at School of Electronic and Information Engineering, Tongji University, Shanghai, China. His research interests include intelligent control, flight control, pattern recognition, and optimal theory.

Yongsen Zheng received the PhD degree in computer science and technology from Sun Yat-sen University, Guangzhou, China in 2023. She is currently a research fellow at College of Computing and Data Science, Nanyang Technological University, Singapore, and is also at Digital Trust Centre, Singapore (DTC). Her current research interests include human-AI dialogue system, conversational recommender system, trustworthy recommendation, natural language processing, trustworthy AI, AI safety, large language models, and causal reasoning.



Leyi Zhao received the BEng degree in computer science from Sichuan University, Chengdu, China in 2021, and the MEng degree in artificial intelligence from University of Manchester, UK in 2023. He is currently a PhD candidate at Indiana University, IN, USA. His research focuses on deep learning and data science, with experience in projects involving Transformer-based music representation learning, sentiment analysis, and image stylization.