Hybrid Sequence Augmentation and Optimized Contrastive Loss Recommendation

Minghui Li, Xiaodong Cai*
School of Information and Communication, Guilin University of Electronic Technology, Guilin, China

***Abstract*—To address the issues of relevance and diversity imbalance in the augmented data and the shortcomings of existing loss functions, this study proposes a recommendation algorithm based on hybrid sequence augmentation and optimized contrastive loss. First, two new data augmentation operators are designed and combined with the existing operators to form a more diversified augmentation strategy. This approach better balances the relevance and diversity of the augmented data, ensuring that the model can make more accurate recommendations when facing various scenarios. Additionally, to optimize the training process of the model, this study also introduces an improved loss function. Unlike the traditional cross-entropy loss, this loss function introduces a temporal accumulation term before calculating the cross-entropy loss, integrating the advantages of binary cross-entropy loss. This overcomes the limitation of traditional methods, which apply cross-entropy loss only at the last timestamp of the sequence, thereby improving the model's accuracy and stability. Experiments on the Beauty, Sports, Yelp, and Home datasets show significant improvements in the Hit@10 and NDCG@10 metrics, demonstrating the effectiveness of the recommendation model based on hybrid sequence augmentation and optimized contrastive loss. Specifically, the Hit metric, which reflects model accuracy, improves by 8.64%, 13.07%, 5.92%, and 19.28% respectively on these four datasets. The NDCG metric, which measures ranking quality, increases by 15.60%, 19.01%, 9.66%, and 20.31% respectively.**

Keywords*—Recommendation algorithm; data sparsity; loss function; sequence augmentation; timestamp optimization

I. INTRODUCTION

Recommendation systems analyze vast amounts of data to help users select items they might be interested in, thereby better meeting their personalized needs. These systems typically make inferences based on users' historical behavior data, preferences, and interests, providing accurate recommendations and saving users the time they would otherwise spend filtering content in an environment of information overload. Sequence recommendation, as a more advanced recommendation technique, predicts future items or content that users may like by analyzing and mining users' historical behavior data within a specific time period. Specifically, sequence recommendation not only focuses on users' historical behavior but also considers the temporal sequence of the behavior, allowing it to more accurately capture users' dynamic changes in interests. For example, the user may have first purchased a phone, then selected headphones, and later became interested in a tablet. Based on this historical data, the recommendation system can predict the user's potential future purchase behavior and infer the user's next possible interest. Sequence recommendation plays a crucial role in various internet applications, especially in scenarios such as e-commerce, video streaming, and social platforms. By deeply mining users' historical behavior and preferences, it helps recommendation systems generate personalized results. However, data sparsity has always been a significant challenge for recommendation systems. Since interaction data between users and items is often scarce, especially in large-scale systems, where many items may have been interacted with by only a few users, it becomes difficult for the system to accurately capture the complex relationships between user preferences and item characteristics. This is especially problematic in sequence recommendation, where models need to handle vast amounts of user behavior data and perform time-series modeling. Due to the sparsity of data, there is often insufficient interaction information between users and items, making it difficult for the model to accurately predict the user's next action, thus affecting the accuracy and effectiveness of the recommendation.

To address the problem of data sparsity, researchers have introduced data augmentation methods. However, there has been limited research on the imbalance between the relevance and diversity of the augmented data, which leads to semantic drift issues or limited performance improvements. In response to this, Dang et al. [1] proposed a new model, BASRec, which designed two new operators, M-Reorder and M-Substitute, and used single-sequence and cross-sequence augmentation modules to solve the above problems. However, previous research has shown that using only Reorder and Substitute operators does not yield the best data augmentation results. Furthermore, BASRec uses the commonly used BCE loss function [2] to calculate contrastive loss. Previous studies have indicated that using the CE loss function [3] in recommendation models may lead to better performance. However, the drawback of CE loss is that it is applied only to the last timestamp of the input sequence, which also affects the model's performance [4].

To address these issues, inspired by the literature [4, 5, 6, 7], this study proposes a recommendation algorithm based on hybrid sequence augmentation and optimized contrastive loss (RM-HSAOCL) on the basis of BASRec. Firstly, to further enhance the effect of data augmentation, this study designed two new data augmentation operators—M-Crop and M-Mask. The M-Crop operator enhances the data by cropping the original data and randomly selecting a subpart of the input sequence, ensuring the diversity of the augmented data while maintaining its relevance to the original data. The M-Mask operator randomly masks part of the data in the input sequence, simulating missing or incomplete information. This operation not only improves the robustness of the model but also helps

*Corresponding Author.

the model better adapt to the issue of incomplete data in real-world scenarios. These two new operators, together with the existing M-Reorder and M-Substitute operators, form a more diversified data augmentation scheme. Through this diversified augmentation strategy, the augmented data not only maintains a certain level of relevance but also greatly improves its diversity, thus enhancing the model's performance in various situations and ensuring more accurate recommendations. In addition to the innovation in data augmentation, this study also designed an improved ICE loss function for optimization. Unlike the traditional Cross-Entropy (CE) loss function, the ICE loss function introduces a time accumulation term before calculating the CE loss, compensating for the limitation of traditional methods, which only apply the CE loss at the final timestamp of the input sequence. As a result, the ICE loss function can capture key information not only at the last moment of the time series but also effectively utilize information from all timestamps in the sequence, improving the model's ability to understand long-term time series data. At the same time, the ICE loss function also incorporates the advantages of BCE loss, optimizing data across all timestamps and further improving the model's performance.

The following outlines the structure of the study: Section II reviews related work. Section III delves into the design of the RM-HSAOCL model, covering hybrid sequence enhancement method, improved loss function and model training loss. Section IV presents and analyzes experimental results to validate the approach. Section V summarizes the work and explores future research directions.

II. RELATED WORK

A. Data Augmentation

Although sequential recommendation models have made significant progress in personalized recommendations, the prevalent issue of data sparsity remains a major bottleneck that limits their performance. In practical applications, many users have limited behavior records, especially for new users or the cold-start problem, where there is often a lack of sufficient interaction data, making it difficult for recommendation systems to effectively capture user preferences. To address this challenge, researchers have proposed data augmentation methods. These methods have shown significant effectiveness in tackling the data sparsity problem in sequential recommendations. By generating more user behavior sequences, introducing generative models, utilizing multimodal data, and performing sample reconstruction, recommendation systems can better cope with the challenges posed by sparse data. It is a commonly used technique, especially in deep learning and machine learning, that helps improve the diversity of the dataset, enhance the model's generalization ability, and reduce the risk of overfitting.

Initially, Tang et al. [8] proposed generating new training samples through a sliding window approach to increase the training data for the model. However, since heuristic algorithms like sliding windows rely solely on local information, the augmented data generated by this method may be of lower quality, potentially leading to overfitting or poor training results. Therefore, researchers gradually introduced many data synthesis methods that require training in order to overcome the limitations of traditional augmentation methods and further enhance the model's generalization ability. For example, Li et al. [9] improved recommendation accuracy and personalization by better capturing users' latent interest shifts through the consideration of spatial and temporal factors. Jiang et al. [10] proposed a conditional Generative Adversarial Network (GAN), which generates new data similar to the original data, thereby enabling the recommendation system to leverage more data. Wang et al. [11] innovatively adjusted user behavior sequences from a counterfactual reasoning perspective, specifically by replacing some of the purchased items with unknown items to simulate different behavior scenarios, helping the model better understand users' latent preferences and decision-making processes. Liu et al. [12] adopted a diffusion model for sequence generation and designed two guidance strategies to control the consistency of the generated data with the original data, ensuring that the generated items maintained a high degree of similarity with users' actual interests. Wang et al. [13] improved temporary user recommendations and reduced cold-start problems by leveraging the behavior features of core users. However, although these training-based data synthesis methods can improve the model's performance to some extent, the augmented data they generate may still have inconsistencies in quality compared to the original data.

To address this issue, many researchers have optimized models by generating contrastive samples using self-supervised learning techniques. For example, Xie et al. [5] proposed a self-supervised learning method based on data augmentation, designing three data augmentation operators and combining them with a contrastive learning framework to improve the model's generalization ability. Yao et al. [14] proposed a two-stage augmentation strategy, where the first stage involves masking operations on the embedding layer, and in the second stage, they discard other classification features except for those used in contrastive learning, to learn more refined feature representations. Zhou et al. [15] also employed contrastive learning, enhancing the model's learning ability by maximizing mutual information between attributes. Their study also introduced random masking of attributes and item order techniques, further improving the model's adaptability to data diversity and noise. Liu et al. [16] combined item similarity information with the contrastive learning objective and proposed a novel data augmentation method, which included insertion and replacement operations. Qiu et al. [17] further optimized the training process by constructing contrastive samples, helping the model recognize subtle differences between different categories, thus improving prediction accuracy. Bian et al. [18] conducted two types of representation augmentation to enhance personalized feature representations of users. Dang et al. proposed five types of data operators to expand item sequences based on time intervals, enhancing the accuracy and effectiveness of recommendations by optimizing the sequence order of time series [19, 20].

However, these methods have limited research on the issue of imbalance between the relevance and diversity of augmented data, which can lead to semantic drift or limited performance improvement. To address this issue, this study designs two new data augmentation operators, M-Crop and M-Mask, based on the BASRec model. These operators help the model better balance the relevance and diversity of the augmented data, making the data augmentation process more refined. This ensures that the augmented data retains key information while providing greater diversity, thereby enhancing the algorithm's adaptability and robustness in practical applications.

B. Loss Function

In recent years, with the rapid development of deep learning technology, sequence-based recommendation models based on neural networks have made significant progress in the field of personalized recommendations. These models, by deeply mining the temporal features in user behavior sequences, are able to more accurately predict users' future preferences. Traditional recommendation systems often focus on predictions based on static user data (such as user history, ratings, etc.), while sequence-based recommendation systems further utilize the time information in user behavior, capturing the dynamic changes in user interests. Specifically, these models usually treat user behavior sequences as input, and through the layers of a neural network, they progressively extract users' potential interests and behavior patterns, thereby generating personalized recommendation content and greatly enhancing the accuracy of recommendations and the user experience.

Among many sequence-based recommendation models, Transformer-based models have particularly attracted widespread attention and favor from researchers. The Transformer architecture was originally proposed to solve sequence-to-sequence tasks (such as machine translation), and its self-attention mechanism allows it to effectively capture long-distance dependencies with high computational efficiency. The first models to apply Transformer to sequence recommendation tasks were SASRec [21] and BERT4Rec [22]. SASRec and BERT4Rec were inspired by the success of the GPT [23] and BERT [24] architectures, which made significant breakthroughs in natural language processing (NLP) tasks, and thus their design ideas were transferred to the sequence recommendation field. Although SASRec and BERT4Rec are similar to their original designs in many aspects, they have made adjustments to the training objectives and attention mechanisms. The authors of BERT4Rec believe that the bidirectionality of the model is the main reason for its performance surpassing that of SASRec. However, subsequent research has shown that the key factor behind the performance improvement is actually the difference in the loss functions used by the two models, while other modifications might have an adverse impact [25, 26]. SASRec is trained using binary cross-entropy (BCE) loss [2], with one positive sample and one negative sample, while BERT4Rec uses cross-entropy (CE) loss [3] across the entire project catalog. This highlights the superiority of CE loss over BCE loss in multi-class classification. However, because BCE demonstrates superior scalability when dealing with larger project portfolios, it is often the more preferred choice in real-world applications.

In addition, many researchers have made improvements to the standard cross-entropy loss. For instance, Li et al. [27] adjusted the weights of positive and negative samples, reducing the loss contribution of simple samples (i.e., easy-to-classify samples), thereby shifting the focus of training to harder-to-classify samples. This adjustment helps the model focus more on predicting difficult user behaviors during the training process, thereby improving the accuracy and effectiveness of the recommendations. These improvements demonstrate that, in practical recommendation systems, how to design an appropriate loss function and sample weighting strategy is often more effective in enhancing model performance than purely architectural innovations. Therefore, this study proposes an improvement to the CE loss by combining the advantages of BCE loss to enhance the model's recommendation performance.

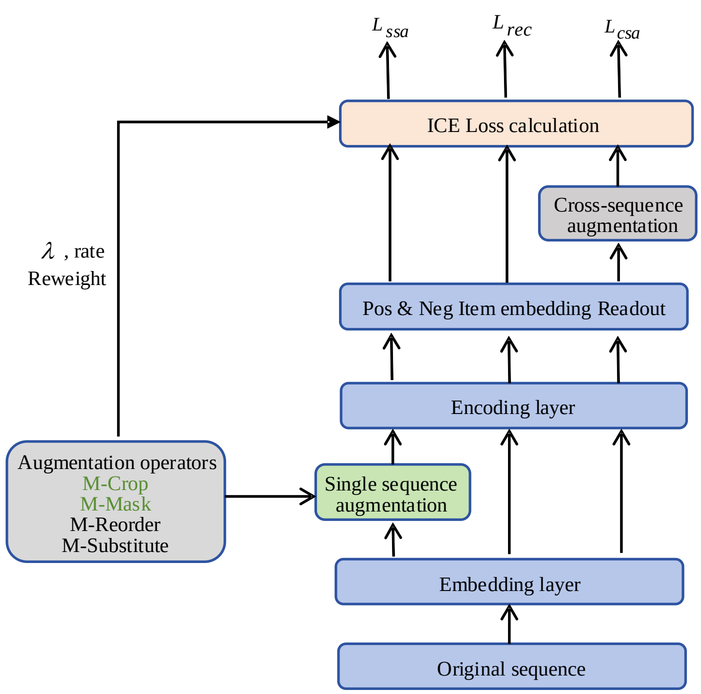

Fig. 1. The Overall framework diagram of the RM-HSAOCL model.

III. RM-HSAOCL MODEL DESIGN

A. Notation, Definition and Description

Sequence recommendation is the task of recommending the next item that a user is likely to interact with based on their historical interaction data. Let U and V denote the set of users and items, respectively. A user $u \in U$ has an interacted item $S_u = \{v_1, v_2, ..., v_j ..., v_N\}$, $v_j \in V(1 \le j \le N)$, denoted as the item interacted by user u at position j in the sequence, where N is the sequence length. Given the historical interactions S_u, the goal of sequence recommendation is to recommend an item from the set of items V that user u is likely to interact with at step N + 1, which can be expressed as formula (1):

$$\arg\max_{v \in V} P(v_{N+1} = v \mid S_u) \tag{1}$$

B. Overall Framework

The overall framework of the RM-HSAOCL model proposed in this study is shown in Fig. 1. After the input sequence passes through the embedding layer, it is processed by the M-Crop and M-Mask operators designed in this section, along with the original M-Reorder and M-Substitute operators, to perform single-sequence enhancement. Then, it goes through the encoding layer, where positive and negative item embeddings are read out, followed by cross-sequence enhancement. Finally, the ICE loss function designed in this section is used to compute the next item prediction loss, single-sequence enhancement loss, and cross-sequence enhancement loss.

C. Hybrid Sequence Enhancement Method

In the BASRec model proposed by Dang et al. [1], two new operators, M-Reorder and M-Substitute, were designed to perform data augmentation. However, previous research shows that using only the Reorder and Substitute operators does not achieve the best data augmentation results. In this section, two new operators, M-Crop and M-Mask, are designed, which, together with the original M-Reorder and M-Substitute operators, complete the data augmentation operation.

1) M-Crop. Random cropping (Crop) is a common and efficient data augmentation technique in computer vision, widely used to improve the generalization ability of deep learning models [5]. Its basic principle is to randomly select a subregion from the original image for cropping and use the cropped image as a training sample. In this way, the model is exposed to different parts of the image, increasing the diversity of the training data, which helps the model make better predictions when faced with new, unseen images. In sequence recommendation tasks, models often need to process long sequence data, which may include user history, click records, or browsing records. To enhance the training data for recommendation algorithms, researchers have introduced the concept of random cropping into this field.

Inspired by the work in [1], this section improves upon the Crop technique and introduces a new data augmentation operator, M-Crop, as shown in Fig. 2. Given an original sequence S_u, M-Crop first selects a subsequence of length $c = rate \cdot N$. This section introduces the method of drawing $rate$ from a uniform distribution, making the length of the augmented subsequence no longer fixed. This allows for the generation of subsequences of varying lengths, thereby increasing the diversity of the augmented data as shown in formula (2):

$$rate \sim Uniform(a, b) \tag{1}$$

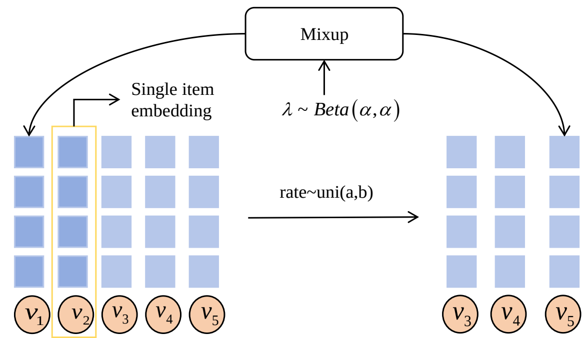

Fig. 2. The Data augmentation operator M-Crop.

where, a and b are hyperparameters, and $0 < a < b < 1$. Then, the augmented sequence is obtained starting from position i as in formula (3):

$$S_u' = Crop(S_u) = [v_i, v_{i+1}, \cdots, v_{i+c-1}] \tag{2}$$

Unlike traditional operators that directly use S_u' as the new sample for model training, this method mixes the corresponding terms of S_u and S_u' together, generating new training samples in the representation space as in formulas (4) and (5):

$$A_u' = Look-up(S_u, S_u') \tag{3}$$

$$A_u^{In} = \lambda \cdot A_u + (1-\lambda) \cdot A_u' \tag{4}$$

where, A_u is the original item representation, $\lambda \sim Beta(\alpha, \alpha)$ is the mixing weight, and A_u^{In} is the augmented representation used for model training.

2) M-Mask. In many natural language processing tasks, such as sentence generation, sentiment analysis, and question answering, the technique of randomly masking input words, also known as "word dropout", is widely used to avoid overfitting. In [5], the authors proposed a data augmentation method called random item masking (Mask). Inspired by the work in [1], this section improves upon Mask and introduces a new data augmentation operator, M-Mask, as shown in Fig. 3. Given the original user sequence S_u, the items $l = rate \cdot N$ in the sequence are masked, where $rate$ is obtained from formula (2), and the augmented sequence can then be obtained as formula (6):

$$S_u{}' = Mask\left(S_u\right) = \left[v_1, v_2, \cdots, v_l, \cdots, v_N\right] \tag{6}$$

By following the same steps as in M-Crop, the augmented representation A_u^{In} can be obtained through formulas (4) and (5).

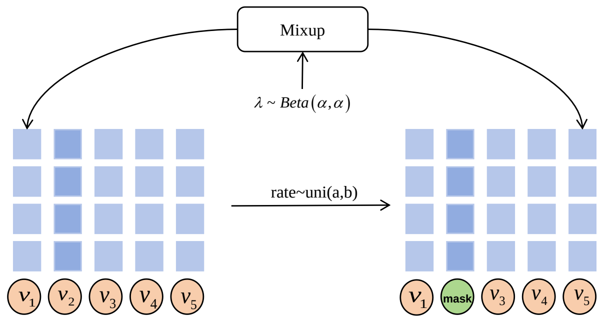

Fig. 3. The Data augmentation operator M-Mask.

D. Improved Loss Function

As discussed in related work, BERT4Rec achieves better performance due to the use of cross-entropy (CE) loss across the entire item sequence [25, 26], whereas in many real-world applications, binary cross-entropy (BCE) loss may be more applicable. The BCE loss is suitable for binary classification problems. Its design allows the model to effectively measure the gap between the predicted values and the true labels, helping the model continuously optimize its predictions during training. The specific mathematical formulas for CE loss and BCE loss are as follows [see formulas (7) and (8)]:

$$CE = -\log \frac{\exp\left(\log i_{t,pos}\right)}{\sum_{c=1}^{C} \exp\left(\log i_{t,c}\right)} \tag{5}$$

$$BCE = -\sum_{t=1}^{l} \left[\log \sigma\left(r_{t,pos}\right) + \sum_{j=1}^{N_s} \log \sigma\left(1 - r_{t,neg_j}\right) \right] \tag{6}$$

Here, l represents the length of the input sequence. The CE loss only involves the final timestamp of the input sequence, as shown in Fig. 4. In contrast, the calculation of BCE loss involves all timestamps of the input sequence, as shown in Fig. 5.

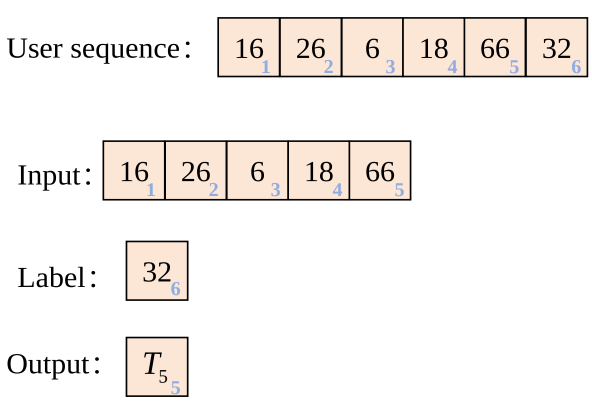

Fig. 4. CE Loss.

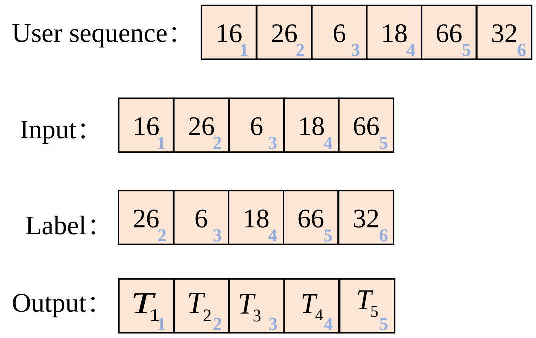

Fig. 5. BCE Loss.

Inspired by the studies [4, 7], this section introduces an optimized loss function, ICE, which builds upon the traditional CE loss function and incorporates an innovative cumulative time term before its calculation. Specifically, the traditional CE loss function usually only considers the last timestamp of the sequence, ignoring the potential information from other timestamps within the sequence. However, in many practical applications, the temporal characteristics of data often have a crucial impact on the model's predictions and performance. To overcome this limitation, the ICE loss function incorporates the accumulation term of time when calculating the CE loss, enabling a more comprehensive consideration of the information at each timestamp in the sequence, thus fully leveraging the temporal dependencies in the data. The specific mathematical formula of the ICE loss function is as follows [formula (9)]:

$$ICE = -\frac{1}{l}\sum_{t=1}^{l} \log \frac{\exp\left(i_{t,pos}\right)}{\sum_{c=1}^{C} \exp\left(i_{t,c}\right)} \tag{7}$$

The core innovation of the ICE loss function lies in the fact that, in the calculation at each timestamp, it takes into account the information from all previous timestamps. This accumulation mechanism effectively captures the long-term dependencies in the input sequence. Specifically, in the traditional CE loss, the model only focuses on the last moment of the sequence, which may overlook potentially useful information in the historical data. In contrast, ICE accumulates the effects of all timestamps, allowing the model to consider the context of the entire time series. This approach enables the model to more accurately capture long-term patterns and trends in the time series, improving the prediction accuracy and reliability.

At the same time, the ICE loss function also draws on the advantages of the BCE loss function, optimizing each timestamp of the entire time series, rather than just the last one. The BCE loss function is typically used in multi-label classification problems. By optimizing the loss at all timestamps, it achieves a more comprehensive optimization, ensuring that the model not only focuses on the final prediction result but also considers the prediction performance at all moments during the process. The ICE loss function combines this idea with the accumulation term of time, gradually optimizing the prediction error at each timestamp during the training process, thereby improving the overall performance of the model.

E. Model Training Loss

This section simultaneously optimizes the entire framework by leveraging the multi-task learning paradigm. The formula for the total model training loss L is as follows [formula (10)]:

$$L = L_{rec} + L_{ssa} + L_{csa} \tag{8}$$

where, L_{rec} is the next item prediction loss, L_{ssa} is the single-sequence enhancement loss, and L_{csa} is the cross-sequence enhancement loss. The calculation methods for these are the same as in BASRec, except that the BCE loss originally used is replaced with the ICE loss function designed in this section, that is [formulas (11),(12),and (13)]:

$$L_{rec} = ICE\left(H_u, A_u^+, A_u^-\right) \tag{9}$$

$$L_{ssa} = \omega \cdot ICE\left(H_u^{In}, A_u^+, A_u^-\right) \tag{10}$$

$$L_{csa} = ICE\left(H_u^{Out}, A_u^{Out+}, A_u^{Out-}\right) \tag{11}$$

F. Model Pseudocode

To help readers understand the workflow of the model, this section provides the pseudocode of the RM-HSAOCL model, as shown in Algorithm 1.

Algorithm 1: Pseudocode of the RM-HSAOCL Model

1: **While** RM-HSAOCL Not Convergence **do**:

2: **for** x in Dataloader (**X**) do:

3: Input the user sequence into the embedding layer for processing;

4: Apply the M-Crop and M-Mask operators designed in this section to the output, along with the original operators, for single-sequence enhancement;

5: Pass through the encoding layer, perform positive and negative item embedding readout, and then apply cross-sequence enhancement;

6: Calculate the next item prediction loss L_{rec} , single-sequence enhancement loss L_{ssa} , and cross-sequence enhancement loss L_{csa} using equation (9), as shown in equations (11) to (13);

7: Calculate the total model training loss L using equations (10);

8: End for

9: End while

10: Return L

IV. Experimental Setup and Results Analysis

A. Experimental Setup

The experimental environment in this section uses Windows 11, with a GPU configuration of RTX 3060 and a memory capacity of 16GB. All the experimental code in this section is written in Python 3.10, and the experimental framework uses PyTorch 1.12.1. All parameter settings are consistent [1].

To ensure the generalizability of the experiment, this section uses four publicly available datasets employed by BASRec: the Beauty, Sports, and Home datasets [1] from the largest e-commerce platform Amazon, as well as the commercial dataset Yelp [2]. For these datasets, the data processing method proposed by Dang et al. [1] is used, which removes users and items with fewer than 5 interactions. Then, a Leave-One-Out strategy is employed to evaluate the performance of each model. During this process, the data is divided based on the timestamps provided in the dataset into training, validation, and test sets. Specifically, for each user, the last record is selected as the test data, the second-to-last record as the validation data, and the remaining records are used as training data. The detailed information for the four datasets is shown in Table I.

TABLE I. Statistical Information of Experimental Datasets

Dataset	#User	#Item	#Action	Sparsity	Avg. Len.
Beauty	22,363	12,101	198,502	99.92%	8.9
Sports	35,598	18,357	296,337	99.95%	8.3
Yelp	30,431	20,033	316,354	99.95%	10.4
Home	66,519	28,237	551,682	99.97%	8.3

This section uses the commonly used evaluation metrics Hit@10 and NDCG@10 for sequence recommendation. The higher the values of Hit and NDCG, the more accurate the recommendations. A high Hit value typically indicates that the recommendation system can provide more items related to the user's interests, while a high NDCG value suggests that the system is able to accurately rank items according to the user's preferences, ensuring that the most relevant recommendations are prioritized. Therefore, the larger the Hit and NDCG values, the higher the accuracy of the recommendation system, leading to a better recommendation experience for users. The system's personalization and precision are also significantly improved.

B. Comparison and Analysis of Model Results

To validate the effectiveness of the RM-HSAOCL model, this section compares it with several representative sequence-based recommendation models: SASRec (2018) [21] introduces the self-attention mechanism into sequence recommendation tasks, addressing the issues of information loss and computational efficiency encountered by traditional sequence models when dealing with long sequences. The self-attention mechanism allows the model to dynamically assign different weights to the user's historical behavior when generating each recommendation, thereby capturing changes in user interest at different time points. ASReP (2021) [10] is a

[1] https://cseweb.ucsd.edu/~jmcauley/

[2] https://www.yelp.com/dataset

pre-training method designed to solve short-sequence recommendation problems. Its core idea is to generate pseudo-items (i.e., pseudo-predictive items) by reversing the input sequence and inserting them at the beginning of the sequence, thereby extending the sequence length. DiffuASR (2023) [12] is a diffusion-model-based sequential recommendation algorithm aimed at solving data sparsity and long-tail user problems in sequential recommendation systems. Its core idea is to use data augmentation techniques and employ diffusion models to generate high-quality pseudo-sequence data, thus enhancing the performance of the recommendation system. CL4SRec (2022) [5] is a sequence recommendation method based on contrastive learning, which addresses data sparsity and improves user representation quality by extracting self-supervised signals from both the original and augmented data using random data augmentation and contrastive learning. BASRec (2025) [1] proposed two new operators, M-Reorder and M-Substitute, for single-sequence augmentation. These operators mix the representations of items in the original sequence with those in the augmented sequence to generate new samples. Along with the cross-sequence augmentation module it designed, these operators perform augmentation and fusion operations to generate new samples that balance relevance and diversity.

Experiments were conducted on the Beauty, Sports, Yelp, and Home datasets, and the results are shown in Table II, where boldface indicates the best performance and underlined text indicates the second-best performance. All improvements are statistically significant, as determined by a paired t-test with the second best result in each case ($p \leqslant 0.05$). From these results, it can be observed that:

Compared to SASRec, ASReP and DiffuASR perform significantly better, which proves that introducing data augmentation methods indeed has a positive effect on mitigating data sparsity, helping to improve the model's robustness and generalization ability. Meanwhile, CL4SRec outperforms ASReP and DiffuASR, indicating that introducing a similarity contrast between augmented data and original data can effectively maintain consistency and quality between the augmented and original data, further enhancing the model's performance. This method ensures that the augmented data's quality is comparable to the original data, avoiding the introduction of excessive noise or inconsistent information, thus optimizing the model's prediction accuracy. However, compared to CL4SRec, BASRec's performance is even more superior, proving that balancing relevance and diversity in the process of generating augmented data is crucial for improving the model's performance. BASRec, by optimizing the augmented data, ensures both the relevance between data and effectively enhances the diversity of the data. This balance not only improves the model's adaptability to different user needs but also enhances its performance in complex recommendation tasks. Therefore, BASRec, by introducing the balance of relevance and diversity in the data augmentation process, effectively overcomes the potential overfitting and information overload issues found in traditional methods, demonstrating outstanding performance across various evaluation metrics.

TABLE II. COMPARISON OF EXPERIMENTAL RESULTS OF VARIOUS MODELS

Dataset	**Metrics**	**SASRec**	**ASReP**	**DiffuASR**	**CL4SRec**	**BASRec**	**Ours**	**Improvement（%）**
Beauty	*Hit@10*	0.0639	0.0664	0.0679	0.0686	<u>0.0810</u>	**0.0880**	8.64
	NDCG@10	0.0338	0.0351	0.0372	0.0366	<u>0.0455</u>	**0.0526**	15.60
Sports	*Hit@10*	0.0320	0.0353	0.0387	0.0412	<u>0.0436</u>	**0.0493**	13.07
	NDCG@10	0.0174	0.0195	0.0202	0.0221	<u>0.0242</u>	**0.0288**	19.01
Yelp	*Hit@10*	0.0277	0.0319	0.0308	<u>0.0355</u>	0.0326	**0.0376**	5.92
	NDCG@10	0.0136	0.0162	0.0150	<u>0.0176</u>	0.0164	**0.0193**	9.66
Home	*Hit@10*	0.0149	0.0184	0.0179	0.0212	<u>0.0223</u>	**0.0266**	19.28
	NDCG@10	0.0078	0.0099	0.0105	0.0119	<u>0.0128</u>	**0.0154**	20.31

The RM-HSAOCL model proposed in this section demonstrates improvements compared to the models mentioned above. The RM-HSAOCL model innovatively introduces two entirely new operators, M-Crop and M-Mask, which, along with the M-Reorder and M-Substitute operators designed in BASRec, participate in the data augmentation process to effectively generate more representative augmented data. This augmentation method not only increases the diversity of the training data but also ensures the quality and relevance of the augmented data, further optimizing the model’s training process. Specifically, the M-Crop operator simulates user behavior changes by randomly cropping different parts of the data, while M-Mask enhances the model's robustness to missing information by randomly masking parts of the data. These operations effectively alleviate the data sparsity problem and improve the model's prediction capability for unseen data. In addition, the RM-HSAOCL model also designs a new loss function, ICE, which adds a time accumulation term before the CE loss, addressing the limitation of applying CE only to the last timestamp of the input sequence. It leverages the advantage of BCE by optimizing across all timestamps. The use of the ICE loss in the joint training loss calculation further enhances the model’s performance. On the Beauty dataset, Hit@10 increased by 8.64%, and NDCG@10 increased by 15.60%. On the Sports dataset, Hit@10 improved by 13.07%, and NDCG@10 increased by 19.01%. On the Yelp dataset, Hit@10 increased by 5.92%, and NDCG@10 increased by 9.66%. On the Home dataset, Hit@10 improved by 19.28%, and NDCG@10 increased by 20.31%. These results show a significant improvement in the recommendation accuracy and ranking quality of the model on these datasets. Overall, the experimental results across all datasets demonstrate substantial improvements on various metrics, indicating that the method is highly applicable and effective in recommendation tasks across different domains.

C. Ablation Experiment

1) Verify the effectiveness of the M-Crop and M-Mask data augmentation operators. In order to analyze the effects of M-Crop and M-Mask, this section designed three variant models using the control variable method: RMHSAOCL-C, which removes the M-Crop operator; RMHSAOCL-M, which removes the M-Mask operator; and RMHSAOCL-CM, which removes both the M-Crop and M-Mask operators. The experiments were conducted on the Beauty, Sports, and Yelp datasets, and the results are shown in Table III.

From the experimental results, it can be observed that on the Beauty and Yelp datasets, the performance metrics of RMHSAOCL-C and RMHSAOCL-M are significantly higher than those of RMHSAOCL-CM. This indicates that adding the M-Crop and M-Mask operators designed in this section can provide more effective data augmentation for the model, which helps improve its performance. Compared to the three variant models, RMHSAOCL achieves the best performance. This result suggests that the synergistic effect of the M-Crop and M-Mask operators significantly enhances the model's learning ability on these two datasets, especially in terms of data augmentation. The generated samples not only exhibit greater diversity but also maintain a high correlation with the original data, thereby avoiding excessive noise interference.

The experimental results on the Sports dataset, however, were different. The performance metrics of RMHSAOCL-M were significantly higher than those of RMHSAOCL-C and RMHSAOCL-CM, and were comparable to the performance of RMHSAOCL. This suggests that, on the Sports dataset, the M-Crop operator played a key role in improving the model's performance, while the M-Mask operator did not significantly enhance the performance. A possible reason for this could be that the features in the Sports dataset are more reliant on the overall structure and temporal aspects of the data, and the M-Mask operator did not provide enough benefit for this type of data. On the other hand, M-Crop, by effectively selecting local data segments, may have helped the model better capture important local patterns and temporal relationships within the data, thus playing a more critical role in improving performance.

In summary, the experimental results indicate that the sensitivity to M-Crop and M-Mask operators varies across different datasets, which suggests that in practical applications, data augmentation strategies may need to be adapted and adjusted for specific datasets.

TABLE III. THE ABLATION EXPERIMENT TO VALIDATE THE EFFECTIVENESS OF THE M-CROP AND M-MASK OPERATORS

Model	**Beauty**		**Sports**		**Yelp**	
	Hit@10	***NDCG@10***	***Hit@10***	***NDCG@10***	***Hit@10***	***NDCG@10***
RMHSAOCL	0.0880	0.0526	0.0493	0.0288	0.0376	0.0193
RMHSAOCL-C	0.0859	0.0513	0.0474	0.0275	0.0363	0.0182
RMHSAOCL-M	0.086	0.0516	0.0498	0.0286	0.0364	0.0190
RMHSAOCL-CM	0.0844	0.0506	0.0477	0.0274	0.0367	0.0180

TABLE IV. THE ABLATION EXPERIMENT TO VALIDATE THE EFFECTIVENESS OF THE ICE LOSS FUNCTION

Model	**Beauty**		**Sports**		**Yelp**	
	Hit@10	***NDCG@10***	***Hit@10***	***NDCG@10***	***Hit@10***	***NDCG@10***
RMHSAOCL	0.0880	0.0526	0.0493	0.0288	0.0376	0.0193
RMHSAOCL+B	0.0787	0.0450	0.0441	0.0239	0.0337	0.0168
RMHSAOCL+C	0.0686	0.0392	0.0332	0.0187	0.0234	0.0113
RMHSAOCL+F	0.0805	0.0449	0.0421	0.0228	0.0320	0.0159

2) Verify the effectiveness of the ICE loss function: To validate the effect of the ICE loss function in the model, three variant models were designed. RMHSAOCL+B indicates the use of the BCE loss function in the model, RMHSAOCL+C indicates the use of the CE loss function in the model, and RMHSAOCL+F indicates the use of the Focal Loss function in the model. The experiments were conducted on the Beauty, Sports, and Yelp datasets, and the experimental results are shown in Table IV.

The experimental results show that, across the three datasets, the performance of RMHSAOCL+B is significantly better than that of RMHSAOCL+C and RMHSAOCL+F. This result indicates that, compared to the CE loss function, the BCE loss function can simultaneously consider the impact of all timestamps in the input sequence and demonstrate superior performance when handling larger datasets. Specifically, the BCE loss function can optimize the model more stably when the ratio of positive and negative samples is relatively balanced, avoiding the training instability or slow convergence issues that the CE loss function may encounter when dealing with larger datasets. The experimental results also show that, when the ratio of positive and negative samples is balanced, the performance of the Focal Loss function is not significantly better than that of the BCE loss function. This could be because Focal Loss focuses more on distinguishing difficult samples, and when the positive and negative samples are already balanced, the advantages of Focal Loss do not manifest. In contrast, the BCE loss function, due to its simple and effective nature, provides better performance.

The performance of the RMHSAOCL model far exceeds that of the other three variant models, proving the effectiveness of the ICE loss function. The ICE loss function optimizes the BCE loss function across all timestamps and combines the advantages of the CE loss function, significantly improving the model's recommendation accuracy and generalization ability. This demonstrates that integrating information from different timestamps to optimize the CE loss function is a key strategy for enhancing model performance. With this approach, the model not only better captures historical information but also effectively improves its performance in practical recommendation scenarios, especially when handling large-scale datasets, where its advantages are even more pronounced.

D. The Impact of Data Augmentation in Mixing Weights on Model Performance

$\lambda \sim Beta(\alpha,\alpha)$ is a mixing weight parameter used to adjust the mixing ratio between the original item representation and the augmented item representation, thereby generating augmented representations for model training. Specifically, the goal of the mixing weight is to find the optimal balance between the original features and the augmented features, so that the model can better learn and capture potential patterns and relationships in the data. The augmented representation extends the original data by introducing certain variations or noise, enabling the model to have stronger generalization ability. The Beta distribution is chosen to describe this mixing process, with its parameters being a hyperparameter that controls the shape and mixing intensity of the distribution, which is crucial for the model’s learning and final performance. In this study, different values were selected for the experiment, ranging from {0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8}, in order to observe the specific impact of different values on the model's performance. The experiment was conducted on the Beauty dataset, and the results are shown in Fig. 6.

From the experimental results, it can be seen that Hit@10 achieves the best value when a = 0.4, meaning that the model's recommendation accuracy is strongest at this setting, allowing it to better predict user interests. NDCG@10 achieves the best values when a = 0.3 and a = 0.5, where the model performs best in terms of ranking accuracy. At a = 0.4, the performance of NDCG@10 is not as good as at 0.3 and 0.5, but still achieves a suboptimal value, indicating that the model at this setting balances recommendation accuracy and ranking performance, maintaining good overall performance.

In summary, different values of the hyperparameter have varied impacts on the model's performance across different metrics. On the Beauty dataset, for the Hit@10 and NDCG@10 metrics, setting a = 0.4 yields good results.

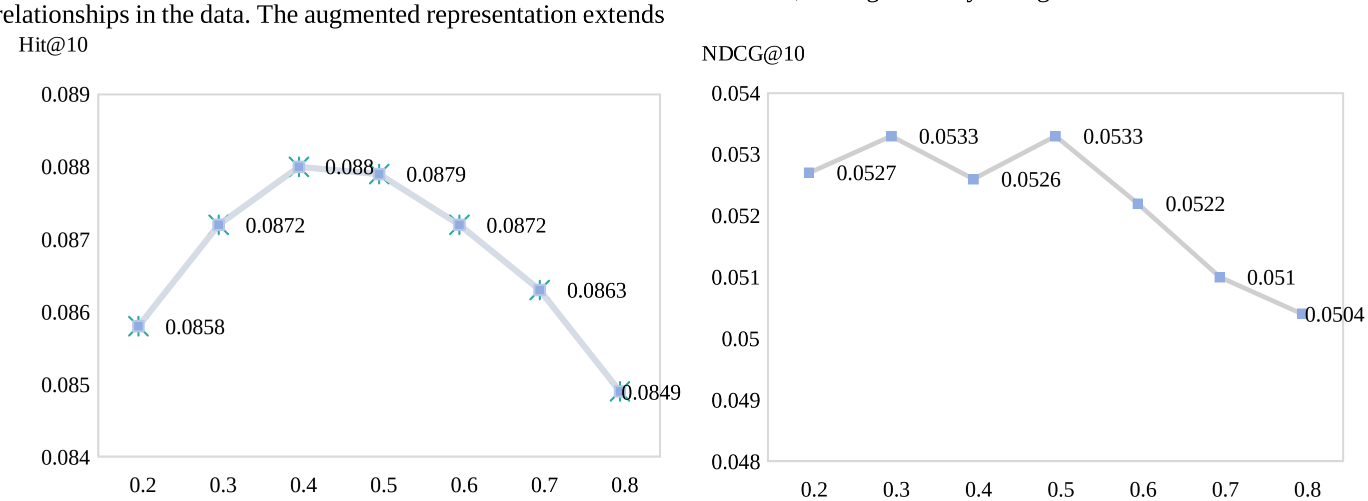

Fig. 6. Sensitivity analysis of the hyperparameter.

V. Conclusion

This study proposes a recommendation algorithm based on mixed sequence augmentation and optimized contrastive loss. By introducing two new data augmentation operators, M-Crop and M-Mask, the data expansion process is enhanced. Additionally, a loss function, ICE, which improves upon the CE loss, is designed for next-item prediction loss, single-sequence augmentation loss, and cross-sequence augmentation loss calculation, thereby improving the accuracy of the recommendation system. The experiments in this study were validated on four public datasets: Beauty, Sports, Yelp, and Home, with significant improvements in various metrics. The application demonstrates notable effects, proving the effectiveness and advancement of the proposed recommendation algorithm based on mixed sequence augmentation and optimized contrastive loss in current sequence recommendation models. Although the algorithm proposed in this study effectively improves the model's performance, in real-world recommendation systems, data may come in various types, such as text, images, videos, etc. The method presented in this study mainly focuses on modeling and augmenting behavior sequence data. However, it may not be effective in integrating and utilizing other types of data, such as user reviews of products or the image features of products. This could result in suboptimal performance when the model deals with multimodal data, as it may fail to fully leverage the information from various data sources to enhance recommendation performance.

Future research can focus on the following aspects to further optimize and expand the algorithm proposed in this study:

Diversity and adaptability of augmentation methods: The current M-Crop and M-Mask operators mainly enhance sequence data. In the future, more augmentation methods targeting different types of data can be explored, especially in recommendation systems that integrate multimodal data. How to design more refined and adaptive augmentation operators will be an important direction.

Interpretability of sequence recommendation models: As recommendation algorithms are increasingly applied in real-world scenarios, the interpretability of models becomes particularly important. Future research can combine the current augmentation and optimization methods to explore how to improve the interpretability of recommendation systems, allowing users to better understand the recommendation logic of the model and increasing their trust in the recommendation system.

Combination of reinforcement learning and transfer learning: Reinforcement learning and transfer learning are technologies that have gradually gained attention in recommendation systems in recent years. Future research could consider combining reinforcement learning with hybrid sequence augmentation methods to dynamically adjust recommendation strategies and optimize users' long-term satisfaction. Transfer learning, on the other hand, can help transfer knowledge from different domains or tasks to the target recommendation task, thereby enhancing the system's cross-domain application ability.

In conclusion, the recommendation algorithm based on hybrid sequence augmentation and optimized contrastive loss proposed in this study has demonstrated superior performance in experiments, with strong practical value. However, as the scale of data grows and recommendation scenarios become more complex, how to further improve the accuracy, efficiency, and scalability of recommendation systems remains an area worthy of in-depth research. Future research can focus on areas such as multimodal recommendation, personalized optimization, and real-time processing, driving recommendation technologies towards more efficient and intelligent development.

References

[1] Dang Y, Zhang J, Liu Y, et al. Augmenting Sequential Recommendation with Balanced Relevance and Diversity[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(11): 11563-11571.

[2] Amjadi M, Mohseni Taheri S D, Tulabandhula T. Katrec: Knowledge aware attentive sequential recommendations[C]//Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings 24. Springer International Publishing, 2021: 305-320.

[3] Alley E C, Khimulya G, Biswas S, et al. Unified rational protein engineering with sequence-based deep representation learning[J]. Nature methods, 2019, 16(12): 1315-1322.

[4] Mezentsev G, Gusak D, Oseledets I, et al. Scalable cross-entropy loss for sequential recommendations with large item catalogs[C]//Proceedings of the 18th ACM Conference on Recommender Systems. 2024: 475-485.

[5] Xie X, Sun F, Liu Z, et al. Contrastive learning for sequential recommendation[C]//2022 IEEE 38th international conference on data engineering (ICDE). IEEE, 2022: 1259-1273.

[6] Liu Z, Chen Y, Li J, et al. Contrastive self-supervised sequential recommendation with robust augmentation[J]. arxiv preprint arxiv:2108.06479, 2021.

[7] Petrov A, Macdonald C. A systematic review and replicability study of bert4rec for sequential recommendation[C]//Proceedings of the 16th ACM Conference on Recommender Systems. 2022: 436-447.

[8] Tang J, Wang K. Personalized top-n sequential recommendation via convolutional sequence embedding[C]//Proceedings of the eleventh ACM international conference on web search and data mining. 2018: 565-573.

[9] Li Y, Luo Y, Zhang Z, et al. Context-aware attention-based data augmentation for POI recommendation[C]//2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW). IEEE, 2019: 177-184.

[10] Jiang J, Luo Y, Kim J B, et al. Sequential recommendation with bidirectional chronological augmentation of transformer[J]. arxiv preprint arxiv:2112.06460, 2021, 90.

[11] Wang Z, Zhang J, Xu H, et al. Counterfactual data-augmented sequential recommendation[C]//Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021: 347-356.

[12] Liu Q, Yan F, Zhao X, et al. Diffusion augmentation for sequential recommendation[C]//Proceedings of the 32nd ACM International conference on information and knowledge management. 2023: 1576-1586.

[13] Wang J, Le Y, Chang B, et al. Learning to augment for casual user recommendation[C]//Proceedings of the ACM Web Conference 2022. 2022: 2183-2194.

[14] Yao T, Yi X, Cheng D Z, et al. Self-supervised learning for large-scale item recommendations[C]//Proceedings of the 30th ACM international conference on information & knowledge management. 2021: 4321-4330.

[15] Zhou K, Wang H, Zhao W X, et al. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization[C]//Proceedings of the 29th ACM international conference on information & knowledge management. 2020: 1893-1902.

[16] Liu Z, Fan Z, Wang Y, et al. Augmenting sequential recommendation with pseudo-prior items via reversely pre-training transformer[C]//Proceedings of the 44th international ACM SIGIR conference on Research and development in information retrieval. 2021: 1608-1612.

[17] Qiu R, Huang Z, Yin H, et al. Contrastive learning for representation degeneration problem in sequential recommendation[C]//Proceedings of the fifteenth ACM international conference on web search and data mining. 2022: 813-823.

[18] Bian S, Zhao W X, Wang J, et al. A relevant and diverse retrieval-enhanced data augmentation framework for sequential recommendation[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022: 2923-2932.

[19] Dang Y, Yang E, Guo G, et al. TiCoSeRec: Augmenting data to uniform sequences by time intervals for effective recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(6): 2686-2700.

[20] Dang Y, Yang E, Guo G, et al. Uniform sequence better: Time interval aware data augmentation for sequential recommendation[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(4): 4225-4232.

[21] Kang W C, McAuley J. Self-attentive sequential recommendation[C]//2018 IEEE international conference on data mining (ICDM). IEEE, 2018: 197-206.

[22] Sun F, Liu J, Wu J, et al. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer[C]//Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 1441-1450.

[23] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.

[24] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019: 4171-4186.

[25] Klenitskiy A, Vasilev A. Turning dross into gold loss: is bert4rec really better than sasrec?[C]//Proceedings of the 17th ACM Conference on Recommender Systems. 2023: 1120-1125.

[26] Petrov A V, Macdonald C. gsasrec: Reducing overconfidence in sequential recommendation trained with negative sampling[C]//Proceedings of the 17th ACM Conference on Recommender Systems. 2023: 116-128.

[27] Li B, Yao Y, Tan J, et al. Equalized focal loss for dense long-tailed object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 6990-6999.