Enhancing Sequential Recommendation with Global Diffusion

Mingxuan Luo[1], Yang Li[2], Chen Lin[1]*,

[1]School of Informatics,Xiamen University,Xiamen,China
[2]Institute of Artificial Intelligence,Xiamen University,Xiamen,China
chenlin@xmu.edu.cn

Abstract

Existing sequential recommendation models are mostly based on sequential models, which can be misled by inconsistent items in the local sequence. This study proposes GlobalDiff, a plug-and-play framework to enhance the performance of sequential models by utilizing a diffusion model to restore the global non-sequential data structure of the item universe and compensate for the local sequential context. Several novel techniques are proposed, including training construction, guided reverse approximator, and inference ensemble, to seamlessly integrate the diffusion model with the sequential model. Extensive experiments on various datasets demonstrate that GlobalDiff can enhance advanced sequential models by an average improvement of 9.67%.

1 Introduction

Sequential Recommendation (SR) aims to predict the next items a user prefers based on the interaction sequence. By leveraging sequential patterns in historical interactions, SR provides more accurate and timely recommendations and enhances user satisfaction and engagement (Kang and McAuley 2018; Sun et al. 2019; Yang et al. 2024).

Most existing SR methods rely on *sequential models*, such as Markov Chain (Rendle, Freudenthaler, and Schmidt-Thieme 2010), GRU (Tan, Xu, and Liu 2016), and Transformer (Kang and McAuley 2018), to capture the temporal dependencies in the interaction sequence. Recent studies point out the *non-sequential item selection problem*, i.e., items in an interaction sequence may not follow strict sequential assumptions (Wang et al. 2017, 2018; Sun et al. 2019). Thus, unidirectional models from left to right are insufficient for learning the representation of an interaction sequence. Bidirectional Transformer encoder has become the state-of-the-art, which fulfills a cloze task in the training phase, i.e., masks a random item in the interaction sequence and predicts the masked item based on the surrounding context (Sun et al. 2019; Du et al. 2022).

Unfortunately, the sequential model is limited in handling *item selection inconsistent with personal interest* due to external factors (Schnabel and Bennett 2020; Zhang et al. 2021). For example, Figure 1(a) depicts Alice, who loves horror movies, selecting 'Mad Max' between 'Puppet Master ' and 'Puppet Master III'. The selection of 'Mad Max' is affected by external factors such as social media or movie promotion and is not driven by Alice's personal interest. The presence of 'Mad Max' in the context window misleads CBiT (Du et al. 2022) to predict 'Mad Max II' incorrectly[1].

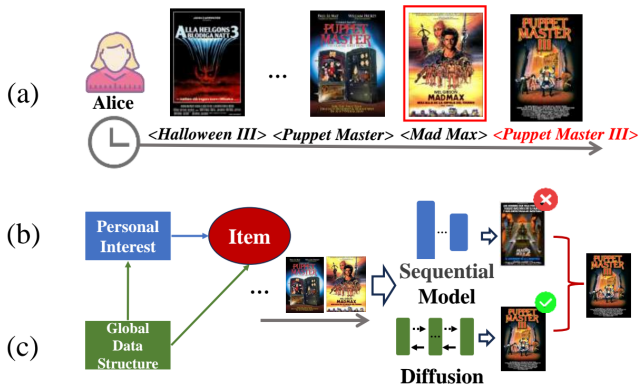

Figure 1: Illustration of Sequential Recommendation: (a) inconsistent item selection in a sequence; (b) traditional assumption that item selection solely relies on personal interest, encoded by the sequential model; (c) our assumption that item is generated based on personal interest and global data structure, which naturally leads to a framework stacking a diffusion model on the sequential model.

The illustration in Figure 1(b) reveals the limitation from the model viewpoint, i.e., sequential models assume the item selection is conditioned only on personal interest. Thus, they encode personal interest from the context window by discriminating an observed item with negative samples. However, from a generative perspective, we argue that there are two types of priors: personal interest and external factors. The external factors are not observable in the local interaction sequence, but they can be inferred from the non-sequential, global data structure, e.g., which item is popular, which items correlate with each other, etc.

As shown in Figure 1 (c), instead of purely relying on the sequential models to encode personal interest in the interac-

*Corresponding Author chenlin@xmu.edu.cn

[1]Example obtained by CBiT on the MovieLens dataset

tion sequence, our goal is to use a separate diffusion model to restore the global non-sequential data structure of the item universe. There are two significant advantages to developing such a paradigm. First, the diffusion model can be plugged and played upon any pre-trained sequential models and enhance *next-item prediction* by balancing the influences of personal interest with the global data structure. Second, the generative nature of the diffusion model allows it to synthesize interaction sequences, mitigate the impact of negative samples in the discriminating objective of sequential models, and increase the accuracy of *long-term prediction*.

Although diffusion models are successfully applied in image generation (Sohl-Dickstein et al. 2015; Liang et al. 2018; Ma et al. 2019), they are infeasible in integration with a sequential model unless they address the following three challenges. **C1 Training sample construction**. Diffusion models operate on continuous signals such as images. Adapting diffusion models to discrete samples, such as items in the interaction sequence, remains under-explored. Using the item embedding (Yang et al. 2024) can not distinguish identical items in different sequences and timestamps, which will remarkably reduce the sample diversity and impede the training efficiency. **C2 Training stability**. Diffusion models often require extensive computational resources and encounter mode collapse during training (Jin et al. 2020; Wang et al. 2017). **C3 Inference ensemble**. On the one hand, different sequential models adopt various architectures. On the other hand, diffusion models are not inherently designed for sequence generation tasks; the diffusion output can not be directly combined with the sequential model.

To address these challenges, we propose GlobalDiff, a plug-and-play framework for sequential recommendation by stacking a diffusion model upon a sequential model.

To address C1, GlobalDiff introduces a novel approach that transforms a masked item within a sequence into a vectorized sample. This sample encapsulates both the item's global associations with other items and its position within the sequence. The transformation scheme generates a diversified training set, making it possible for GlobalDiff to learn patterns beneficial for the SR task (Section 3.3).

To address C2, GlobalDiff includes a standard forward process to add Gaussian noise gradually to the original sample, and a *guided reverse approximator* to reconstruct the original sample in the reverse process. The reverse approximator can accommodate any pre-trained sequential model and utilize the semantic context information learned by the sequential model. Thus, GlobalDiff can lead to more personalized items in the sequence and tackle the training instability issues (Section 3.4).

To address C3, GlobalDiff proposes a novel inference ensemble method. Despite the different architectures, all sequential models have the same form of output, i.e., a softmax distribution, which estimates personal preference over items. The diffusion model reconstructs the global preference over items from the sequential model's output. GlobalDiff adopts a voting mechanism to combine personal and global preferences and reduces the risk of over-fitting to abnormal item selection in a sequence (Section 3.5).

Experimental results on various datasets show that GlobalDiff can improve the next item prediction of advanced SR models by 9.67% across all datasets and evaluation metrics. GlobalDiff improves the SOTA CBiT (Du et al. 2022) by an average of 5.17% across all evaluation metrics, e.g., HR and NDCG. GlobalDiff also improves the long-term prediction accuracy, i.e., HR@20 for the next four items, by an average improvement of 9.40% across three backbone models and all datasets.

To sum up, the contributions of this work are as follows.

- We propose a plug-and-play framework GlobalDiff to enhance the recommendation performance of sequential models by stacking a diffusion model without any changes to the original sequential model.
- We propose a series of novel techniques, including training construction, guided reverse approximator, and inference ensemble, to seamlessly integrate the diffusion model with the sequential model.
- We conduct substantial experiments on three common benchmark datasets to validate that GlobalDiff can improve the next item prediction and long-term prediction accuracy over advanced backbone sequential models.

2 Related Work

Sequential recommendation (SR). Conventional recommendation (Rendle et al. 2012) ignores the order of actions and assumes that all user-item historical interactions are equally important. On the contrary, Sequential recommendation (Boka, Niu, and Neupane 2024) and session-based recommendation (Wang et al. 2021) analyze the temporal dependencies of user-item interactions to improve the accuracy of predictions and have received numerous attention from both academia and industry (Kang and McAuley 2018; Xie et al. 2022; Wang et al. 2020). To enhance the performance of SR, various neural network architectures have been integrated into the framework, including CNN (Tang and Wang 2018), RNN (Liu and Singh 2016; Rath and Sahu 2020) and the GRU variant (Hidasi et al. 2015), and Transformer (de Souza Pereira Moreira et al. 2021; Sun et al. 2019; Kang and McAuley 2018; Du et al. 2022). Specifically, Unidirectional Transformers, which is used in SASRec (Kang and McAuley 2018), model sequences in a left-to-right manner, focusing on past interactions. Bidirectional Transformers, such as those seen in BERT4Rec and CBiT (Sun et al. 2019; Du et al. 2022), consider both past and future interactions, providing a more comprehensive context window for understanding user behavior.

Diffusion Models (DM). The diffusion model is an emerging research area in domains such as image generation (Ho, Jain, and Abbeel 2020; Song and Ermon 2020) and Natural Language Processing (NLP) (Li et al. 2022; Gong et al. 2022). Early applications of DM are concentrated on the conventional recommendation task, which directly implements the diffusion process on the user-item implicit feedback matrix (Wang et al. 2023; Walker et al. 2022). These models are inferior for sequential recommendation as they have difficulties handling sequence data. Diffusion models for sequential recommendation mostly model the interacted items in a sequence, i.e., the forward process

perturbs the representation of an interacted item by adding Gaussian noise, and the reverse process reconstructs the item representation (Li, Sun, and Li 2023; Lee and Kim 2024; Du et al. 2023; Yang et al. 2024). Some recent studies also use the diffusion model to generate pseudo interactions for data augmentation (Ma et al. 2024; Liu et al. 2023).

Remarks. Our work (GlobalDiff) significantly differs from the existing diffusion models for sequential recommendations. First, instead of using the diffusion model as the main predictor (Li, Sun, and Li 2023; Lee and Kim 2024; Du et al. 2023; Yang et al. 2024), GlobalDiff integrates with a backbone sequential model in a plug-and-play manner. Therefore, GlobalDiff is more flexible and can achieve superior performance with more advanced sequential models. Second, unlike DiffuASR (Liu et al. 2023) and PDRec (Ma et al. 2024), which claim themselves to be a plug-in module on sequential models, GlobalDiff does not change the training of the backbone sequential model. Therefore, GlobalDiff is less intrusive and requires significantly less computational resources, as GlobalDiff can be directly built upon a pre-trained backbone sequential model.

3 Methodology

3.1 Preliminaries

Given a set of users $\mathcal{U}$ and a set of items $\mathcal{I}$, where the number of users is $|\mathcal{U}| = N$ and the number of items is $|\mathcal{I}| = M$. Suppose the interaction history is a set of sequences $s^u \in \mathcal{S}$ where each user u has an interaction sequence $s^u = [s_1^u, s_2^u, ..., s_{L_u}^u]$, and $L_u \leq L$ is the length of the sequence, with L being the maximal sequence length. Most Sequential Recommendation (SR) methods forecast the next item in the sequence $s_{L_u+1}^u$. In this paper, we also consider the long-term item prediction, i.e., $s_{L_u+P}^u, P > 1$, given the training set of interaction history $\mathcal{S}$.

As shown in Fig.2, GlobalDiff stacks a diffusion model upon a backbone sequential model. First, the backbone model is trained prior to GlobalDiff with its own objectives (Section 3.2). The parameters of the backbone model are fixed in the training stage of GlobalDiff. Next, the training samples are transformed from $\mathcal{S}$ for the diffusion model (Section 3.3). The diffusion model consists of a forward (noising) process and a reverse (denoising) process, while an approximator in the reverse process is trained to recover a disrupted sample obtained by the forward process (Section 3.4). Finally, after GlobalDiff has finished training, the predictions of the diffusion model and sequential model are combined in the inference stage (Section 3.5).

3.2 Backbone Model

Without loss of generality, the SR backbone model contains an *encoder* and a *projection* module. The *encoder* reads a sequence $s^u \in \mathcal{S}$ and derives the representation at each position $\mathbf{h}^u = (\mathbf{h}_1^u, \cdots, \mathbf{h}_{L_u}^u)$. Assuming the target position is i, the *projection* module maps the representation $\mathbf{h}^u$ to a distribution over the item universe:

$$\mathbf{y}^{u,i} = softmax(g(\mathbf{h}^u)), \tag{1}$$

where $\mathbf{y}^{u,i} \in \mathbb{R}^M$ is the probability distribution over possible items, $g(\cdot)$ is the projection function. The projection function is defined by the original design of the backbone model. For instance, in CBiT (Du et al. 2022), a linear layer is used for projection e.g., $g(\mathbf{h}^u) = \mathbf{W} \cdot \mathbf{h}_i^u + \mathbf{b}$, where $\mathbf{W}$ is the weight matrix and $\mathbf{b}$ is the bias term.

The most common objective a backbone model attempts to minimize is the cross entropy loss of the prediction of target item $\mathbf{y}^{u,i}$ versus the ground truth s_i^u. A unidirectional model such as SASRec (Kang and McAuley 2018) implements the *next-item-prediction* task, i.e., predicting $\mathbf{y}^{u,i}$ based on $[s_1^u, \cdots, s_{i-1}^u]$. Bidirectional models such as Bert4Rec and CBiT (Sun et al. 2019; Du et al. 2022) implement the *cloze* task, i.e., masking a random position i and predicting $\mathbf{y}^{u,i}$ based on $[s_1^u, \cdots, s_{L^u}^u]$. We do not change the training task and objective of the backbone model.

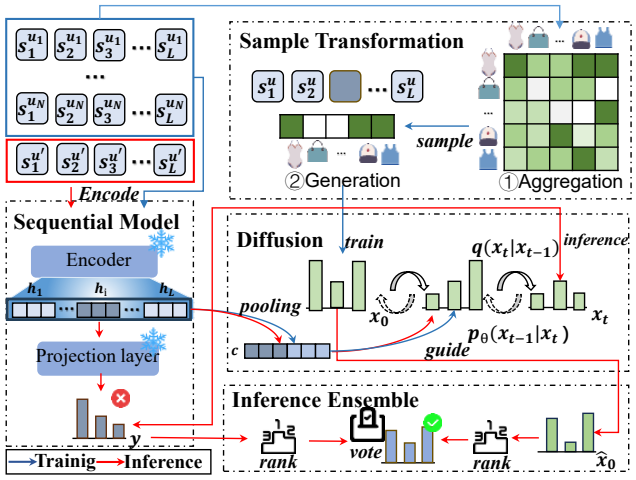

Figure 2: Framework of GlobalDiff: blue flow indicates the training process, red flow indicates the inference process.

3.3 Training Sample Transformation

As shown in Figure 2, the workflow of training sample transformation for GlobalDiff consists of two major steps.

Step 1: Aggregation. As introduced in Section 1, inconsistent item selection may disturb the learning of personal interest from an individual sequence. We believe that introducing the global non-sequential item structure can mitigate this issue.

To do so, we first aggregate historic interaction sequences across users to obtain a binary user-item interaction matrix $\mathbf{O} \in \mathbb{R}^{N \times M}$, where $\mathbf{O}_{u,i} = 1$ indicates item i has appeared in the interaction sequence s^u.

Then, we compute the item-item similarity matrix $\mathbf{A} \in \mathbb{R}^{M \times M}$ based on the user-item interaction matrix $\mathbf{O}$. Each cell measures the item similarity, i.e., $\mathbf{A}_{i,j} = cos(\mathbf{O}_{:,i}, \mathbf{O}_{:,j})$, where $cos(\cdot,\cdot)$ is the cosine similarity, and $\mathbf{O}_{:,j}$ is the $j-$th column of the user-item interaction matrix. Thus, items with similar user behaviors will have a higher similarity in $\mathbf{A}$.

Step 2: Generation. We derive the training samples based on *masking*. We are motivated to adopt the masking scheme for two reasons. First, masking and the cloze task are widely adopted in advanced SR methods (Sun et al. 2019; Du et al.

2022). If the backbone model adopts the masking training scheme, then the training set for the diffusion model aligns with the backbone sequential model, which can help us to control and balance the training of the two models. Second, masking has shown superiority in enabling models to understand the context containing non-sequential item selection (Sun et al. 2019), i.e., item selections are not strictly sequential. If the backbone model does not adopt the masking scheme, then the training set for the diffusion model can compensate with the backbone model and reduce the non-sequential item selection problem.

Formally, an item is randomly masked for each sequence $s^u \in \mathcal{S}$. Suppose the $i-$th item is masked, i.e., $\hat{s}^{u,i} = (s_1^u, \cdots, [\text{mask}], \cdots, s_{L_u}^u)$. The masking is repeated on each sequence several times to construct a training dataset $\hat{\mathcal{S}} = \{\hat{s}^{u,i}\}$.

Diffusion models can not be fed with discrete items directly. Thus, for each masked sequence $\hat{s}^{u,i} \in \hat{\mathcal{S}}$, the goal of this step is to generate a vector $\tilde{\mathbf{s}}^{u,i} \in \mathbb{R}^M$, which represents the masked item s_i^u by its globally alike items.

First, we obtain the corresponding row in the item-item similarity matrix, i.e, $\mathbf{A}_{s_i^u,:}$, and initiate a vector $\mathbf{q}^{u,i} \in \mathbb{R}^M$ for each masked sequence $\hat{s}^{u,i}$.

$$\mathbf{q}_m^{u,i} = \mathbf{A}_{s_i^u,m}, \forall 1 \leq m \leq M, \tag{2}$$

where $\mathbf{q}_m^{u,i}$ is the $m-$th element in $\mathbf{q}^{u,i}$, which by definition corresponds to the correlation between the masked item s_i^u and item m.

$\mathbf{A}_{s_i^u,:}$ does not distinguish the same item in different sequences, i.e., if $\exists u, u', i, j$ and the masked items are the same $s_i^u = s_j^{u'}$, then we will have the same row extracted from $\mathbf{A}$. We take into account the temporal information within a sequence to adjust some elements in $\mathbf{q}^{u,i}$.

$$\mathbf{q}_{s_j^u}^{u,i} = \min\left[(\frac{\zeta}{|i-j|} + \delta) \cdot \mathbf{A}_{s_i^u,s_j^u}, 1\right], \forall 1 \leq j \leq L_u, s_i^u \neq s_j^u. \tag{3}$$

For any item s_j^u contained in the sequence and s_j^u is not the masked item, i.e., $s_i^u \neq s_j^u$, then the item-item similarity is adjusted by the position difference between i and j. ζ is a scale factor controlling the influence of the relative position on the similarity, with a default value of 1.5. δ is an additional constant term used to adjust the re-weighting, with a default value of 1. We use the cut-off function $\min[\cdot, 1]$ to preserve that $\mathbf{q}_m^{u,i} \in [0, 1]$.

Since $\mathbf{q}_m^{u,i}$ can be interpreted as a probability function, we use Bernoulli sampling on each cell of $\mathbf{q}^{u,i}$ to generate $\tilde{\mathbf{s}}^{u,i}$, i.e., $p(\tilde{\mathbf{s}}_m^{u,i} = 1) = \mathbf{q}_m^{u,i}$. The sampling procedure is designed for two reasons: (1) it can distinguish two different sequences with the same item in the same position, and (2) it can add randomness to increase the sample diversity.

3.4 Difussion Model

The training sample $\tilde{\mathbf{s}}^{u,i}$, which represents a masked item in the context of a sequence based on globally associated items, is fed into the diffusion model. The diffusion model contains two processes: the forward process and the reverse process.

Let's denote the initial sample as $x^{u,i,0} = \tilde{\mathbf{s}}^{u,i}$. The **forward process** progressively adds Gaussian noise to disrupt the initial sample. According to the original derivation of diffusion model (Ho, Jain, and Abbeel 2020), we can directly derive $x^{u,i,T}$ at step T as:

$$x^{u,i,T} = \sqrt{\bar{\alpha_T}} x^{u,i,0} + \sqrt{1 - \bar{\alpha_T}}\epsilon, \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, \mathcal{I})$ is Gaussian noise. Here, according to the derivation from DDPM (Ho, Jain, and Abbeel 2020), $\alpha_t = 1 - \beta_t, \bar{\alpha_T} = \prod_{t=0}^{T} \alpha_t$ and $[\beta_1, \beta_2, ...\beta_T]$ is a set of variables that vary over time.

In the **reverse process**, the diffusion model restores the initial sample by progressively deriving from the disrupted sample at step T through intermediate states $x^{u,i,t}, t \leq T$. The reverse process can be computed as follows:

$$\mu_{t-1}(x^{u,i,t}, x^{u,i,0}) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha_t}} x^{u,i,t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha_t}} x^{u,i,0}, \tag{5}$$

where μ_{t-1} is used to estimate the distribution mean.

An approximator, e.g., a deep neural network(U-Net), is usually employed to estimate $x^{u,i,0}$. Previous study (Wang et al. 2023) identifies that a straightforward application of Equation 5 often results in non-personalized generated items due to the lack of guidance from historical interactions. Thus, as shown in Figure 2, GlobalDiff uses the backbone model's contextual information to enhance the diffusion model's ability to reconstruct the initial sample. Formally,

$$\begin{aligned} \mu_{t-1}(x^{u,i,t}, x^{u,i,0}) = &\frac{\sqrt{\alpha_t}(1 - \bar{\alpha_{t-1}})}{1 - \bar{\alpha_t}} x^{u,i,t} \\ &+ \frac{\sqrt{\bar{\alpha_{t-1}}}\beta_t}{1 - \bar{\alpha_t}} f_\theta(x^{u,i,t}, t, \mathbf{c}^{u,i}), \\ \mathbf{c}^{u,i} = &\left[\mathbf{h}_i^u, \frac{1}{L_u}\sum_{j=1}^{L_u} \mathbf{h}_j^u\right]. \end{aligned} \tag{6}$$

where $f_\theta(\cdot)$ is a U-Net to predict $x^{u,i,0}$ based on $x^{u,i,t}$, t, and $\mathbf{c}^{u,i}$. To fully exploit the semantic representation learned by the sequential model, $\mathbf{c}^{u,i}$ concatenates the representation of the masked item $\mathbf{h}_i^u$, and the mean pooling of representation vectors for all items in the encoder's output $\mathbf{h}^u$. It is important to point out that although some diffusion models (Du et al. 2023; Yang et al. 2024) also use the contextual information in a sequence to guide the reverse approximator, their contextual information is co-learned with the approximator. On the contrary, we use a fixed sequential model, and its extracted contextual information speeds up the learning of the diffusion model and increases the learning stability.

The goal of the general diffusion model is to align the posterior distribution in the forward process with the prior distribution in the reverse process, which can be formulated as a KL divergence. It can be resorted to the mean-squared error loss as follows:

$$\mathcal{L} = \sum_{u,i,t} \mathbb{E}_q[\|f_\theta(x^{u,i,t}, i, \mathbf{c}^{u,i}) - x^{u,i,0}\|_2^2]. \tag{7}$$

3.5 Inference Ensemble

As shown in Figure 2, in inference time, each testing sequence $s^{u'}$ goes through the backbone sequential model, and

the output of the sequential model is $\mathbf{y}^{u',L_{u'}+1}$ by Equation 1. The output is a decision based on the local sequential information, and we are motivated to balance it with the global non-sequential information. Ideally, the global information can be represented by the $s^{u'}_{L_{u'}+1}$'s globally alike items. However, since in inference time, the item to be predicted is unknown, we can not fetch the corresponding row in $\mathbf{A}$ to construct $\mathbf{q}^{u',L_{u'}+1}$.

As illustrated in Figure 1 (c), the personal interest is influenced by the global data structure. Thus, the output of the sequential model denoted as $\mathbf{y}^{u',L_{u'}+1}$, which intends to capture personal interest, can be seen as a disrupted observation of $\mathbf{q}^{u',L_{u'}+1}$, i.e., an intermediate step in the forward noising process. Our idea is to utilize diffusion model to restore $\mathbf{q}^{u',L_{u'}+1}$ from $\mathbf{y}^{u',L_{u'}+1}$. Specifically, for the prediction distribution $\mathbf{y}^{u',L_{u'}+1}$, let $\hat{\mathbf{x}}^{T'} = \mathbf{y}^{u',L_{u'}+1}, T' \leq T$, where T is the maximal diffusion model step in Section 3.4, continue the forward process to obtain $\hat{\mathbf{x}}^T$. Then, initiate a reverse denoising process on $\hat{\mathbf{x}}^T$ to obtain $\hat{\mathbf{x}}^0$. $\hat{\mathbf{x}}^0$ can be seen as an estimated $\mathbf{q}^{u',L_{u'}+1}$, i.e., representation of an item in the context of a sequence based on global non-sequential information.

To align with probability distribution $\mathbf{y}^{u',L_{u'}+1}$, we normalize $\hat{\mathbf{x}}^0$. Then, we adopt the Borda Count $(Saari\ 2012)$, known as a positional voting rule, to combine $\mathbf{y}^{u',L_{u'}+1}, \hat{\mathbf{x}}^0$.

$$score(i) = \gamma \cdot r(\mathbf{y}_i^{u',L_{u'}+1}) \cdot \mathbf{y}_i^{u',L_{u'}+1} + (1-\gamma) \cdot r(\hat{\mathbf{x}}_i^0) \cdot \hat{\mathbf{x}}_i^0, \quad (8)$$

where $i \in \mathcal{I}$ and the $score(i)$ is the final output score for item i. The function $r(\cdot)$ represents the ranking. For example, the highest ranked item in $\mathbf{y}^{u',L_{u'}+1}$ receives a rank of M, and the second item receives a rank of $M-1$, and so on. γ is a weight parameter that balances the contribution of both the backbone model and the diffusion model. In our work, γ is set to the default value of 0.5. We use a greedy search method in inference, i.e., the item with the highest score $score(i)$ is selected as the prediction output.

4 Experiments

In this section, we study the following research questions:

RQ1: Can GlobalDiff improve the performance of the backbone model in next-item prediction?

RQ2: Can GlobalDiff enhance the long-term prediction accuracy of the backbone model?

RQ3: How does each component in GlobalDiff contribute to its performance?

4.1 Experimental Setup

Datasets. We conduct experiments on three publicly available datasets: MovieLens 1M (ML-1M)[2], Amazon-Beauty[3], and KuaiRec[4]. These datasets are commonly adopted to evaluate sequential recommendations (Yang et al. 2024; Sun et al. 2019; Kang and McAuley 2018). They cover different domains and demonstrate various user behavior patterns. For example, MovieLens 1M comprises ratings on movies, Amazon-Beauty captures user interactions on beauty products in an E-commerce platform, and KuaiRec contains interactions from a short-video platform. The statistics of datasets are reported in Table 1.

Dataset	#Users	#Items	#Interations	#Avg_len	#Sparsity
ML-1M	5,180	3,526	562,800	108.60	96.90%
Beauty	1,308	9,708	24,742	18.90	99.80%
KuaiRec	7,176	10,596	8,459,425	1,178.80	88.80%

Table 1: Statistics of datasets after pre-processing

Pre-processing. Following established methods (Sun et al. 2019; Tan et al. 2021), we pre-process the data by treating all ratings as implicit feedback organized chronologically by their timestamps. Unpopular items and users with few interactions are filtered out.

Implementation. We implement all models with Python 3.8 and PyTorch 2.0.1. As for the sequential recommendation model, we fix the maximum sequence length as 10 for all three datasets. In the training stage, the number of diffusion steps $T = 20$. The scale factor $\zeta = 1.5$, re-weighting term $\delta = 1$, and the score weighing coefficient $\gamma = 0.5$. To optimize GlobalDiff, we employ the Adam optimizer, setting the batch size and learning rate to 256 and 0.001. Other hyper-parameters of the three backbone models are set to the default values as mentioned in their original paper. Our codes are available online [5].

Evaluation Metrics. We evaluate the Sequential recommendation performance of all models on three datasets using two widely-used metrics, NDCG@K and Recall@K, where K=[10,20].

Backbones. We implement GlobalDiff on three advanced backbone sequential models. (1) **SRGNN** (Wu et al. 2019) leverages graph neural networks for capturing complex item dependencies. (2) **SASRec** (Kang and McAuley 2018) utilizes self-attention mechanisms. (3) **CBiT** (Du et al. 2022) employs bidirectional Transformer architectures for contrastive learning in sequential recommendation tasks. (4) **BERT4Rec** (Sun et al. 2019) adapts bidirectional Transformer models pre-trained on large-scale interaction data for sequential recommendation tasks.

Competitors. We compare GlobalDiff to several recent SR methods, including (1) **Gru4Rec** (Hidasi et al. 2015), a GRU network is adopted as the encoder to capture temporal dependencies in the interaction sequence. (2) **Caser** (Tang and Wang 2018) uses convolutional neural networks to capture user behavior dependencies. (3) **LightSANS** (Fan et al. 2021) introduces low-rank decomposed self-attention to condense historical items into latent interests for context-aware representations (4) **DuoRec** (Qiu et al. 2022) tackles clustering issues in item embeddings using contrastive regularization, model-level Dropout augmentation, and a novel sampling strategy. (5) **CL4Rec** (Xie et al. 2022) integrates next-item prediction with contrastive learning to enhance self-supervision from user behavior sequences. (6)

[2]https://grouplens.org/datasets/movielens/

[3]https://jmcauley.ucsd.edu/data/amazon/

[4]https://kuairec.com/

[5]https://github.com/XMUDM/GlobalDiff

AC_TSR (Zhou et al. 2023) improves attention weight accuracy in transformer-based models by introducing calibrators. (7) **DiffuRec** (Li, Sun, and Li 2023) , a pioneering use of diffusion models in sequential recommendation.

4.2 Next-Item Prediction

To answer RQ1, we adopt a leave-one-out strategy for evaluation, the most recent interaction is used for testing, the second-to-last for validation, and the rest for training. The experimental results are summarized in Table 2. We have the following observations.

(1) Compared with the vanilla backbone models, *GlobalDiff consistently improves the backbone models in terms of all evaluation metrics and across all datasets*. The average improvement is 9.67%. The improvement indicates that GlobalDiff effectively mitigates the influence of inconsistent items and leads to more accurate recommendations.

(2) Across all backbone models, *GlobalDiff shows the greatest improvement on SasRec*, with an average performance gain of 17.06%. SasRec, being a unidirectional model, struggles with contextual understanding, but GlobalDiff 's use of cloze training samples helps mitigate this issue. Furthermore, even though CBiT achieves state-of-the-art (SOTA) performance compared with other baselines, *incorporating GlobalDiff results in a 5.10% average performance gain on CBiT across all datasets*. This demonstrates that leveraging global item associations with GlobalDiff enhances performance beyond the capabilities of sequence-based learning.

(3) Among all datasets, *GlobalDiff demonstrates the most pronounced improvement on the Beauty dataset*, with an average gain of 13.06%, while the gain on KuaiRec is more modest at 5.53%. This variation is attributed to dataset sparsity. As shown in Table 1, the Beauty dataset is the sparsest, and the interaction sequences are shorter. It is challenging for traditional models to capture strong patterns due to insufficient signals and insufficient contextual information. GlobalDiff addresses this by employing a diffusion process that enhances item-item relationship modeling, effectively mitigating the impact of dataset sparsity.

(4) Among the competitors, DiffuRec's performance is worse than that of the backbone models BERT4Rec and CBiT. This observation indicates that using the diffusion model as the main predictor is suboptimal, and it is necessary to combine the diffusion model with an advanced sequential model.

4.3 Long-term Predictions

To address RQ2, we partition each sequence of length L_u into three segments, the segment of positions $1, \cdots, L_u - 5$ for training, the last four items at positions $L_u - 3, \cdots, L_u$ for testing, and the one item at position $L_u - 4$ for validation. Because in long-term prediction, there is a larger search space of item candidates, we report HR@20 and NDCG@20 results in predicting the item at position $l + P, P = 1, 2, 3, 4$ in figure3, when l is the length of the training sequence. We have the following observations.

(1) In general, GlobalDiff enhances the backbone models across different time steps P, achieving an average improvement of 9.40% in HR@20 and 8.09% in NDCG@20. There is a trend of declining performance with increasing P, which shows that sequential models are not proficient in making long-term predictions. However, even at a longer period, i.e., $P = 4$, GlobalDiff continues to deliver an average performance boost of 6.30% across three backbones. This evidence underscores that non-sequential information captured by GlobalDiff can correct the long-term sequential prediction.

(2) Among the three backbones, GlobalDiff shows the most consistent performance improvement for CBiT. From step 1 to step 4, average gains in HR@20 increase from 12.39% to 12.45%, and in NDCG@20 from 9.52% to 13.20%. This stable enhancement can be attributed to CBiT's use of cloze tasks and contrastive learning on the same sequence samples, which bolsters the robustness of guide representations. As a result, GlobalDiff's effectiveness is significantly reinforced, leading to robust performance gains.

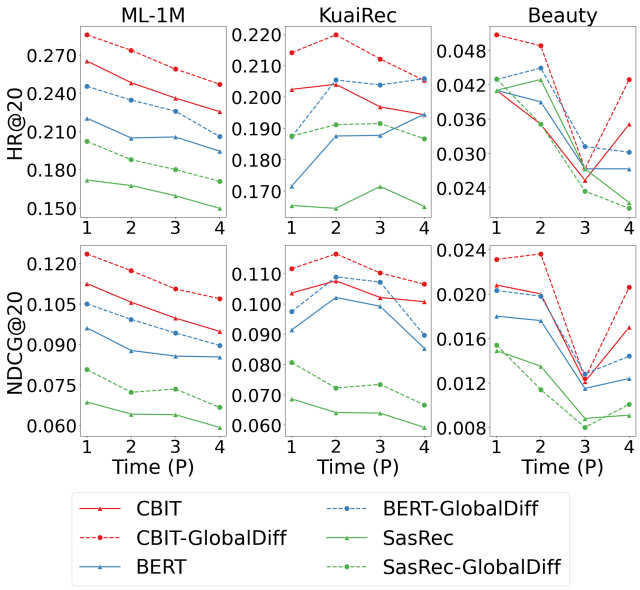

Figure 3: Long-term predictions on the three Datasets

4.4 Ablation Study

To answer RQ3, we conduct the following ablation experiments. We experiment with the most competitive backbone model CBiT on the most widely used dataset, ML-1M.

Ablation on different reverse approximators In Equation 6, the context representation $\mathbf{c}^{u,i}$ is incorporated to guide the training of the reverse approximator. We implement variants of the reverse approximator, (1) with guided representation, which is Equation 6, (2) with only targeted representation, which removes the mean pooling of $\frac{1}{L_u}\sum_{j=1}^{L_u}\mathbf{h}_j^u$ and keeping $\mathbf{h}_i^u$, and (3) without guided representation, which removes $\mathbf{c}^{u,i}$ in Equation 6. We perform next-item prediction to evaluate different approximations. The results are shown in Figure 4.

Methods	Ml-1M				KuaiRec				Beauty			
	HR@10	HR@20	NDCG@10	NDCG@20	HR@10	HR@20	NDCG@10	NDCG@20	HR@10	HR@20	NDCG@10	NDCG@20
Gru4Rec	0.0990	0.1689	0.0469	0.0644	0.1419	0.1941	0.0872	0.1004	0.0222	0.0306	0.0118	0.0140
Caser	0.0485	0.0944	0.0224	0.0340	0.1172	0.1714	0.0686	0.0823	0.0183	0.0252	0.0084	0.0102
LightSANS	0.1023	0.1793	0.0477	0.0670	0.1378	0.1957	0.0804	0.0969	0.0245	0.0352	0.0134	0.0160
DuoRec	0.1276	0.1975	0.0587	0.0762	0.1265	0.1860	0.0731	0.0881	0.0237	0.0352	0.0125	0.0153
CL4Rec	0.0936	0.1523	0.0419	0.0567	0.1129	0.1639	0.0653	0.0781	0.0099	0.0206	0.0048	0.0075
AC-TSR	0.1295	0.2075	0.0667	0.0816	0.1482	0.1924	0.0868	0.1032	0.0222	0.0398	0.0107	0.0151
DiffuRec	0.1320	0.2028	0.0672	0.0851	0.1447	0.1957	0.0876	0.1004	0.0229	0.0321	0.0124	0.0147
SRGNN	0.1134	0.1882	0.0566	0.0754	0.1318	0.1852	0.0794	0.0928	0.0234	0.0367	0.0144	0.0177
+GlobalDiff	0.1158	0.1869	0.0575	0.0754	0.1402	0.1886	0.0850	0.0972	0.0281	0.0390	0.0161	0.0188
Improv.(%)	2.12%	-0.69%	1.57%	0.00%	6.35%	1.81%	7.03%	4.79%	20.09%	6.27%	11.81%	6.21%
SASRec	0.1002	0.1761	0.0506	0.0696	0.1302	0.1844	0.0781	0.0918	0.0226	0.0375	0.0092	0.0129
+GlobalDiff	0.1271	0.1976	0.0617	0.0794	0.1467	0.1981	0.0907	0.1035	0.0289	0.0437	0.0115	0.0152
Improv.(%)	26.85%	12.21%	21.94%	14.08%	12.67%	7.43%	16.13%	12.75%	27.88%	16.53%	25.00%	17.83%
BERT4Rec	0.1283	0.1888	0.0679	0.0831	0.1442	0.1911	0.0912	0.1029	0.0375	0.0476	0.0207	0.0233
+GlobalDiff	0.1398	0.2171	0.0728	0.0923	0.1537	0.2017	0.0964	0.1085	0.0453	0.0578	0.0222	0.0253
Improv.(%)	8.90%	14.90%	7.22%	11.07%	6.59%	5.55%	5.70%	5.44%	20.80%	21.43%	7.25%	8.58%
CBiT	0.1542	0.2375	0.0824	0.1033	0.1590	0.2080	0.0982	0.1105	0.0476	0.0585	0.0284	0.0312
+GlobalDiff	**0.1638**	**0.2503**	**0.0870**	**0.1088**	**0.1672**	**0.2190**	**0.1029**	**0.1160**	**0.0510**	**0.0625**	**0.0292**	**0.0320**
Improv.(%)	6.23%	5.39%	5.58%	5.32%	5.16%	5.29%	4.79%	4.98%	7.14%	6.84%	2.82%	2.56%

Table 2: Experimental results on the three datasets. The best results are in boldface, and the second-best underlined.

(1) Incorporating the contextual representation of backbone models to guide the reverse approximator enhances the predictions of GlobalDiff. The removal of the guided representation causes GlobalDiff to neglect sequential item dependencies, resulting in the poorest performance. Compared to using only the Target Representation, without guided representation leads to an average performance decrease of 81%. The performance degradation is partially attributed to the training instability of the diffusion model to recover the item in a sequence with only global information.

(2) Incorporating the sequence's representation can enhance GlobalDiff's ability. As opposed to using only the target item representation, adding the mean pooling of all items yields a 30% average performance improvement. This enhancement results from richer contextual information and user behavior patterns, thereby significantly improving the diffusion model's training effectiveness.

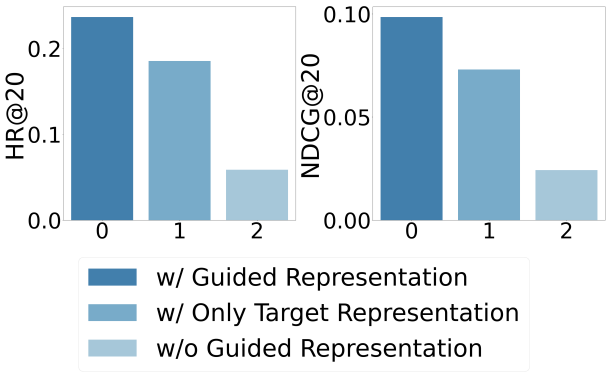

Figure 4: Performance w.r.t variants of reverse approximator

Ablation on inference ensemble To examine the impact of voting in inference, we pause the training of the diffusion model in GlobalDiff after every five training epochs and let the sequential model, the diffusion model, and the ensemble make predictions on the testing set in Section 4.2. The HR@20 and NDCG@20 results are shown in Figure 5.

Even when training is insufficient (i.e., with fewer epochs), the diffusion model can complement the sequential model. For example, after 10 epochs of early-stage training, the diffusion model achieves an HR@20 of 0.168, which is significantly lower than the HR@20 performance of the pretrained sequential model (0.237). However, after applying voting integration, GlobalDiff' HR@20 is 0.247, increasing the sequential model by 4.2%. This improvement is due to the diffusion model's ability to predict items that the sequential model fails to identify accurately. This observation validates the necessity of integrating the diffusion model with the sequential model, as the diffusion model, even at an early stage, can be advantageous.

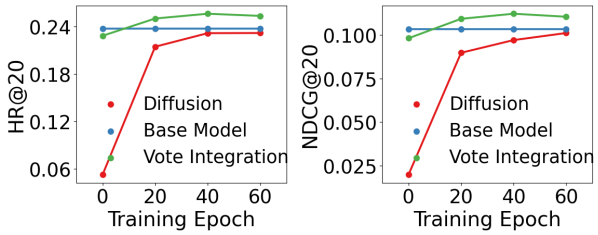

Figure 5: Voting performance with different training epochs

Impact of diffusion steps The training overhead of GlobalDiff increases with the number of diffusion steps T. To assess the impact of T, we report the HR@20 performance of GlobalDiff with $T = 20, 100, 200$ in Figure 6.

We observe that the influence of T is insignificant throughout the training process. Regardless of the number of training epochs, the variation in HR@20 with different diffusion steps is less than 0.005. This observation indicates that GlobalDiff achieves competitive performance with fewer diffusion steps, leading to a substantial reduction in both training and inference costs.

4.5 Discussion on Time Complexity

Finally, we analyze the time complexity of GlobalDiff. GlobalDiff contains an aggregation process with a complex-

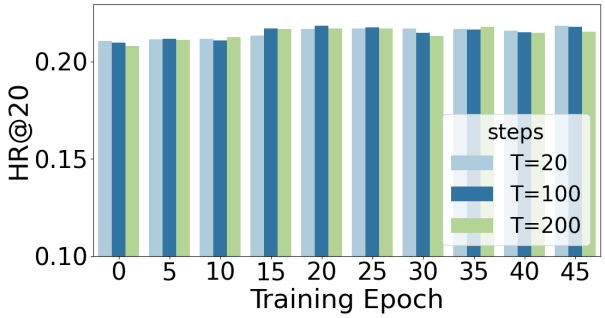

Figure 6: Impact of diffusion steps

ity of $O(M^2)$, where M is the number of items due to pair-wise item-item similarity computation in Section 3.3. When time complexity is a concern, exact pair-wise similarity computation, though possible offline, is unnecessary as the similarity is converted into a probability function for generating training samples. Thus, we can adopt a simple alternative by computing cluster-item similarity, drastically reducing the time complexity from $O(M^2)$ to $O(KM), K < M$, where K is the number of clusters. Specifically, we first cluster items into K clusters using K-means. Then, we compute the item-cluster similarity matrix, i.e.,$\mathbf{A}_{i,j}$ is the similarity between item i and cluster j. In the sample generation steps, we use the item-cluster similarity to generate training samples.

We validate the alternative on CBiT using the Beauty dataset. The results in Table 3 demonstrate that (1) when the number of clusters reaches a certain value, e.g., K=1000, we can achieve comparable performance to exact pair-wise similarity. (2) When K is relatively small, e.g., K=500, the aggregation CBiT+GlobalDiff still outperforms the vanilla CBIT. These results indicate that cluster-item similarity offers a scalable solution without sacrificing performance.

Clusters	HR@10	HR@20	NDCG@10	NDCG@20
N/A (CBiT)	0.0476	0.0585	0.0284	0.0312
K=100	0.0480	0.0580	0.0283	0.0311
K=500	0.0492	0.0618	0.0286	0.0317
K=1000	0.0500	**0.0625**	**0.0295**	**0.0328**
Pair-wise (M=9708)	**0.0510**	**0.0625**	0.0292	0.0320

Table 3: Performance comparison using cluster-item similarity on the Beauty dataset. The best results are in boldface.

5 Conclusion

This paper addresses a major problem in sequential recommendation, i.e., item selection is not only affected by personal interest but also global factors. A novel plug-and-play framework GlobalDiff is presented to integrate a diffusion model with a sequential model to enhance recommendation performance by introducing the global data structure. GlobalDiff makes efforts in training construction, guided reverse approximator, and inference ensemble to support seamless integration. Experiments across various datasets validate GlobalDiff's ability to improve the performance of an advanced sequential model substantially. In the future, we plan to explore more ways to reduce the complexity of GlobalDiff.

Acknowledgments

Chen Lin is the corresponding author. This work is supported by the Natural Science Foundation of China (No.62372390)

References

Boka, T. F.; Niu, Z.; and Neupane, R. B. 2024. A survey of sequential recommendation systems: Techniques, evaluation, and future directions. *Information Systems*, 102427.

de Souza Pereira Moreira, G.; Rabhi, S.; Lee, J. M.; Ak, R.; and Oldridge, E. 2021. Transformers4rec: Bridging the gap between nlp and sequential/session-based recommendation. In *Proceedings of the 15th ACM conference on recommender systems*, 143–153.

Du, H.; Shi, H.; Zhao, P.; Wang, D.; Sheng, V. S.; Liu, Y.; Liu, G.; and Zhao, L. 2022. Contrastive learning with bidirectional transformers for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 396–405.

Du, H.; Yuan, H.; Huang, Z.; Zhao, P.; and Zhou, X. 2023. Sequential recommendation with diffusion models. *arXiv preprint arXiv:2304.04541*.

Fan, X.; Liu, Z.; Lian, J.; Zhao, W. X.; Xie, X.; and Wen, J.-R. 2021. Lighter and better: low-rank decomposed self-attention networks for next-item recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 1733–1737.

Gong, S.; Li, M.; Feng, J.; Wu, Z.; and Kong, L. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jin, B.; Lian, D.; Liu, Z.; Liu, Q.; Ma, J.; Xie, X.; and Chen, E. 2020. Sampling-decomposable generative adversarial recommender. *Advances in Neural Information Processing Systems*, 33: 22629–22639.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, 197–206. IEEE.

Lee, H.; and Kim, J. 2024. EDiffuRec: An Enhanced Diffusion Model for Sequential Recommendation. *Mathematics*, 12(12): 1795.

Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P. S.; and Hashimoto, T. B. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35: 4328–4343.

Li, Z.; Sun, A.; and Li, C. 2023. Diffurec: A diffusion model for sequential recommendation. *ACM Transactions on Information Systems*, 42(3): 1–28.

Liang, D.; Krishnan, R. G.; Hoffman, M. D.; and Jebara, T. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, 689–698.

Liu, D. Z.; and Singh, G. 2016. A recurrent neural network based recommendation system. In *International Conference on Recent Trends in Engineering, Science & Technology*.

Liu, Q.; Yan, F.; Zhao, X.; Du, Z.; Guo, H.; Tang, R.; and Tian, F. 2023. Diffusion augmentation for sequential recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1576–1586.

Ma, H.; Xie, R.; Meng, L.; Chen, X.; Zhang, X.; Lin, L.; and Kang, Z. 2024. Plug-in diffusion model for sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8886–8894.

Ma, J.; Zhou, C.; Cui, P.; Yang, H.; and Zhu, W. 2019. Learning disentangled representations for recommendation. *Advances in neural information processing systems*, 32.

Qiu, R.; Huang, Z.; Yin, H.; and Wang, Z. 2022. Contrastive learning for representation degeneration problem in sequential recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, 813–823.

Rath, A.; and Sahu, S. R. 2020. Recurrent neural networks for recommender systems. *Computational Intelligence and Machine Learning*, 1(1): 31–36.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Rendle, S.; Freudenthaler, C.; and Schmidt-Thieme, L. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, 811–820.

Saari, D. G. 2012. *Basic geometry of voting*. Springer Science & Business Media.

Schnabel, T.; and Bennett, P. N. 2020. Debiasing item-to-item recommendations with small annotated datasets. In *Proceedings of the 14th ACM Conference on Recommender Systems*, 73–81.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, Y.; and Ermon, S. 2020. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33: 12438–12448.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1441–1450.

Tan, Q.; Zhang, J.; Liu, N.; Huang, X.; Yang, H.; Zhou, J.; and Hu, X. 2021. Dynamic memory based attention network for sequential recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 4384–4392.

Tan, Y. K.; Xu, X.; and Liu, Y. 2016. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st workshop on deep learning for recommender systems*, 17–22.

Tang, J.; and Wang, K. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 565–573.

Walker, J.; Zhong, T.; Zhang, F.; Gao, Q.; and Zhou, F. 2022. Recommendation via collaborative diffusion generative model. In *International Conference on Knowledge Science, Engineering and Management*, 593–605. Springer.

Wang, J.; Yu, L.; Zhang, W.; Gong, Y.; Xu, Y.; Wang, B.; Zhang, P.; and Zhang, D. 2017. Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, 515–524.

Wang, S.; Cao, L.; Wang, Y.; Sheng, Q. Z.; Orgun, M. A.; and Lian, D. 2021. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)*, 54(7): 1–38.

Wang, S.; Hu, L.; Cao, L.; Huang, X.; Lian, D.; and Liu, W. 2018. Attention-based transactional context embedding for next-item recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Wang, W.; Xu, Y.; Feng, F.; Lin, X.; He, X.; and Chua, T.-S. 2023. Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 832–841.

Wang, Z.; Wei, W.; Cong, G.; Li, X.-L.; Mao, X.-L.; and Qiu, M. 2020. Global context enhanced graph neural networks for session-based recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 169–178.

Wu, S.; Tang, Y.; Zhu, Y.; Wang, L.; Xie, X.; and Tan, T. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 346–353.

Xie, X.; Sun, F.; Liu, Z.; Wu, S.; Gao, J.; Zhang, J.; Ding, B.; and Cui, B. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*, 1259–1273. IEEE.

Yang, Z.; Wu, J.; Wang, Z.; Wang, X.; Yuan, Y.; and He, X. 2024. Generate What You Prefer: Reshaping Sequential Recommendation via Guided Diffusion. *Advances in Neural Information Processing Systems*, 36.

Zhang, S.; Yao, D.; Zhao, Z.; Chua, T.-S.; and Wu, F. 2021. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 367–377.

Zhou, P.; Ye, Q.; Xie, Y.; Gao, J.; Wang, S.; Kim, J. B.; You, C.; and Kim, S. 2023. Attention calibration for transformer-based sequential recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3595–3605.