Diff4Rec: Sequential Recommendation with Curriculum-scheduled Diffusion Augmentation

Zihao Wu
Department of Computer Science and Technology, Tsinghua University
wuzh22@mails.tsinghua.edu.cn

Xin Wang*
DCST, Tsinghua University
Key Laboratory of Pervasive Computing, Ministry of Education
xin_wang@tsinghua.edu.cn

Hong Chen
Department of Computer Science and Technology, Tsinghua University
h-chen20@mails.tsinghua.edu.cn

Kaidong Li
Yi Han
Yangshipin Co., Ltd.
likaidong@yangshipin.cn
hanyi@yangshipin.cn

Lifeng Sun
DCST, Tsinghua University
Key Laboratory of Pervasive Computing, Ministry of Education
sunlf@tsinghua.edu.cn

Wenwu Zhu*
DCST, Tsinghua University
Key Laboratory of Pervasive Computing, Ministry of Education
wwzhu@tsinghua.edu.cn

ABSTRACT

Sequential recommender systems often suffer from performance drops due to the data-sparsity issue in real-world scenarios. To address this issue, we bravely take advantage of the strength in diffusion model to conduct data augmentation for sequential recommendation in this paper. However, there remain two critical challenges for this scarcely-explored topic: (i) previous diffusion models are mostly designed for image generation aiming to capture pixel patterns, which can hardly be applied in data augmentation for sequential recommendation aiming to capture the user-item relations; (ii) given a specific diffusion model capable of user-item interaction augmentation, it is non-trivial to guarantee that the diffusion-generated data can always bring benefits towards the sequential recommendation model. To tackle these challenges, we propose Diff4Rec, a curriculum-scheduled diffusion augmentation framework for sequential recommendation. Specifically, a diffusion model is pre-trained on recommendation data via corrupting and reconstructing the user-item interactions in the latent space, and the generated predictions are leveraged to produce diversified augmentations for the sparse user-item interactions. Subsequently, a curriculum scheduling strategy is designed to progressively feed the diffusion-generated samples into the sequential recommenders, with respect to two levels, i.e., *interaction augmentation* and *objective augmentation*, to jointly optimize the data and model. Extensive experiments demonstrate that our proposed Diff4Rec framework is able to effectively achieve superior performance over several strong baselines, capable of making high-quality and robust sequential recommendations. We believe the proposed Diff4Rec has the promising potential to bring paradigm shift in multimedia recommendation.

*Corresponding Authors

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612709

CCS CONCEPTS

• **Information systems** → **Information systems applications**.

KEYWORDS

recommender systems, diffusion models, curriculum learning

ACM Reference Format:
Zihao Wu, Xin Wang, Hong Chen, Kaidong Li, Yi Han, Lifeng Sun, and Wenwu Zhu. 2023. Diff4Rec: Sequential Recommendation with Curriculum-scheduled Diffusion Augmentation . In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3581783.3612709

1 INTRODUCTION

In real-world online web services, sequential actions are pivotal to recommender systems, including activities such as browsing products, clicking links, making comments and favoring music etc. Consequently, there has been a surge of interest in sequential recommendation, which seeks to anticipate the subsequent user behaviors based on sequential user interactions with items. Conventional sequential recommendation models have tried to leverage Markov Chain to model temporal interaction transitions [9, 30, 40]. With the advancement of deep learning, recent studies have made remarkable strides in deep sequential recommendation models, which employ various neural networks as sequence encoders to capture dynamic patterns of user behaviors [11, 12, 15, 20, 21, 35, 36].

Nevertheless, sequential recommenders may suffer from performance drop due to the data-sparsity issue in real-world scenarios [9, 19, 25]. Previous works utilize various strategies such as contrastive learning [39, 44], meta learning [5], disentangled representation learning [26], etc. to mitigate the sparsity issue. Among these methods, data augmentation is known as the most direct and effective way to tackle the data-sparsity issue [2, 8, 43, 46]. Therefore, in this paper, we bravely propose to take advantage of the strength within the advanced diffusion models for user-item interactions augmentation, which is expected to improve sequential recommendation performance. However, this topic is scarcely explored in the literature and poses the following two challenges:

(1) Existing diffusion models are mostly designed for image generation which aims to capture visual pixel patterns, failing to

conduct data augmentation for sequential recommendation. It is non-trivial to incorporate the learning ability of diffusion models into the recommendation domain and generate reliable interactions aligned with user intents.

(2) Given a particular diffusion model for user-item interaction augmentation, it is non-trivial to ensure that the data generated through diffusion process can always benefit sequential recommendation.

To tackle these challenges, we propose Diff4Rec, a novel curriculum-scheduled diffusion augmentation framework which is able to generate user-item interactive data for augmentation in sequential recommenders. In particular, a diffusion model tailored for recommendation is designed to gradually corrupt and recover user-item interactions, which are encoded into a latent space to compress complex information and capture latent user intents. The augmented samples generated by the pre-trained diffusion model are then evaluated via a curriculum learning scheduler to be progressively fed into the sequential recommender. The augmentations are conducted in terms of two levels, i.e., *Interaction Augmentation* and *Objective Augmentation*, to effectively utilize the diffusion-generated samples. *Interaction Augmentation* uses the generated samples to enrich the historical sequences and reveal diversified and undiscovered user intentions, while *Objective Augmentation* leverages the generated samples to serve as candidate items for an augmented training objective, thus resulting in better parameter optimizations and model performances. Additionally, an easy-to-hard curriculum training strategy is proposed to alleviate the potential noises hidden in the diffusion-generated samples. Experiments demonstrate the superiority of our proposed Diff4Rec framework.

To summarize, this work makes the following contributions.

- We propose to study the feasibility of leveraging diffusion models to achieve user-item interactive data augmentation for sequential recommendation.
- We propose the curriculum-scheduled diffusion augmentation framework Diff4Rec, consisting of i) a diffusion model capable of modeling user-item interactions in the latent space and ii) a curriculum scheduler which progressively augments user-item interactive sequences with diffusion-generated samples from *interaction augmentation* and *objective augmentation* levels.
- We conduct extensive experiments on several real-world datasets to demonstrate the superiority of the proposed Diff4Rec, which discovers that it is promising to employ diffusion models to facilitate the performance improvement of sequential recommendation, especially for mitigating the data-sparsity problem.

2 RELATED WORK

Sequential Recommendation. Sequential recommendation suggests items to users based on their previous interactions such as purchases, clicks, or ratings. Deep neural networks, such as GRU [12], CNN [36], and other variants, were utilized to demonstrate a high level of efficacy in accomplishing sequential recommendation. Moreover, some researchers used self-attention mechanisms to model the mutual influence between past interactions and achieved impressive improvements in performance [15, 35, 47]. Graph neural networks are also effective at incorporating high-order relationships in a sequence by propagating and aggregating information, and have been applied to sequential recommendation as well [7, 20]. Alternatively, recent research in sequential recommendation has begun to investigate different training strategies. For example, BERT4Rec [35] employs a Cloze objective that predicts masked items in a sequence based on their context. Ma et al.[26] propose a seq2seq training strategy based on disentanglement that predicts the future sequence rather than the next item.

Diffusion Model. Diffusion models [13, 33, 34] essentially learns a step-wise generator that maps Gaussian distribution to data samples in a denoising manner. For its impressive generation ability, diffusion model has been widely explored in computer vision [23, 28, 31, 32], audio processing [3, 17, 27] and AI for science [14, 24, 45]. Recently, some effors have been attempted to employ diffusion model to recommendation [38, 41]. However, these attempts either suffer from poor performance or fail to leverage the data generation ability of diffusion models to enhance the sequential recommendation.

Curriculum Learning. Curriculum learning is a dynamic sample reweighting learning strategy that usually starts with simple patterns and gradually advances to more complex ones, mirroring how humans learn through a structured curriculum [42]. Bengio et al. [1] investigated curriculum learning and demonstrated through empirical studies that such an approach can reduce training time and sometimes even improve generalization. The main components of curriculum learning include a difficulty metric to identify easy data and a training scheduler to determine when to introduce more challenging data for training. Chen et al. [4] explore easy-to-hard curriculum learning on a meta-learning paradigm to transfer the knowledge from multiple citiesto cold-start cities for next POI recommendation. Bian et al. [2] follow the evaluate-and-schedule process and develop a curriculum learning strategy to conduct contrastive learning for modeling sequential user behaviors.

3 DIFF4REC: THE PROPOSED FRAMEWORK

The overall Diff4Rec framework is shown in Figure 1, where we initially utilize a latent diffusion model to predict user-item interactions in a denoising manner. Afterward, with the predictions generated by the pre-trained diffusion model, we augment the interaction sequences to enhance sample diversity and feed the samples into downstream sequential recommenders optimized with curriculum learning strategy. In this section, we will first describe the preliminaries, then the diffusion training for recommendation, and finally the curriculum scheduler.

3.1 Preliminaries and Notations

3.1.1 Diffusion Model. A diffusion model typically involves forward and reverse processes. To begin, a data point that has been sampled from a real-world data distribution $\mathbf{z}_0 \sim q(\mathbf{z})$ is subjected to the forward process, which gradually corrupts $\mathbf{z}_0$ into a standard Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$. Throughout each forward step $t \in [1, 2, ..., T]$, the perturbation is modulated by $q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I})$, where $\beta_t \in (0, 1)$ represents different variance scales. In the reverse phase, the denoising process is initiated, which aims to gradually reconstruct the original data $\mathbf{z}_0$ by sampling from $\mathbf{z}_T$ and a neural network parameterized by θ.

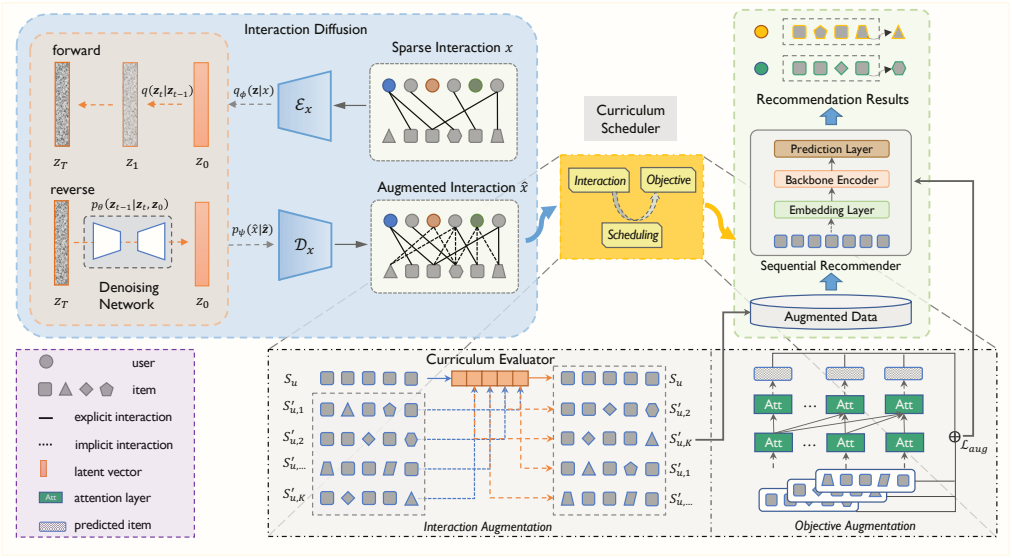

Figure 1: The overall structure of our proposed Diff4Rec framework. We adopt a two-stage paradigm to firstly train a diffusion model on recommendation data, and then learn the sequential recommender through curriculum-scheduling strategy, where the *interaction augmentation* and *objective augmentation* enrich sparse interactions and implement improved optimizations.

3.1.2 Sequential Recommendation. Let $\mathcal{U}$ and $\mathcal{I}$ denote the user and item set, respectively. For each user $u \in \mathcal{U}$, a sequentially ordered interaction sequence $S_N = [i_1, i_2, ..., i_N]$ is provided, where each element $i_j \in \mathcal{I}$ is an item that was interacted, and N is the length of the sequence. The goal of sequential recommendation is to predict a list of items that the user may be interested in, based on its historical sequence S_N leading up to the target time step N.

3.2 Diffusion Training

The diffusion training process aims to obtain a diffusion model that can generate user-item interactions for sequential recommendation, which includes the interaction encoding, diffusion forward, and reverse processes. Next, we describe the three processes in detail.

3.2.1 Interaction Encoding. To lower the computational demands of training diffusion models towards interaction modeling, we propose to utilize a variational autoencoder (VAE) [16] to compress the user-item interactions into latent space whilst making use of the sparse signals to capture the latent intentions and preferences of user behaviors [22, 25]. Given the user set $\mathcal{U}$ and item set $\mathcal{I}$, we denote $x \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$ as the interaction history matrix of users and items, where $x_{u,i}$ represents whether user u has interacted with item i or not. Specifically, a variational encoder $\mathcal{E}$ is utilized to encode user interactions x into a low-dimensional representation $\mathbf{z}$, such that

$$q_\phi(\mathbf{z}|x) = \mathcal{N}\left(\mathbf{z}; \mu_\phi(x), \sigma^2_\phi(x)\mathbf{I}\right), \tag{1}$$

where $\mu_\phi(x)$ and $\sigma^2_\theta(x)\mathbf{I}$ are the mean and variance of the variational distribution, and vector $\mathbf{z}$ will be delivered to the subsequent diffusion model. Correspondingly, given the recovered $\hat{\mathbf{z}}$ from the diffusion model, a decoder $\mathcal{D}$ is utilized to reconstruct user interactions from the latent vector $\hat{\mathbf{z}}$ as $p_\psi(\hat{x}|\hat{\mathbf{z}})$.

Generally, the set of VAE could be optimized by maximizing the variational evidence lower bound (ELBO) [16],

$$\mathcal{L}(\psi, \phi; x, \mathbf{z}) = -D_{KL}\left(q_\phi(\mathbf{z}|x) \| p_\psi(\mathbf{z})\right) + \mathbb{E}_{q_\phi(\mathbf{z}|x)}\left[\log p_\psi(x|\mathbf{z})\right]. \tag{2}$$

With the variational autoencoder consisting of $\mathcal{E}$ and $\mathcal{D}$, we can efficiently access a low-dimensional latent space in which complex user-item interactions are projected into latent vectors. Compared to the high-dimensional interaction data space, the latent space is more appropriate for likelihood-based generative models [31], which enables the neural networks to identify the significant features hidden in the interaction data and perform training in a computationally feasible space.

3.2.2 Forward Process. The forward process gradually corrupts the user interaction vector by adding Gaussian noises, where the transition is parameterized by:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}\left(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}\right), \tag{3}$$

where β stands for a variance schedule, i.e., the scale of Gaussian noise added at each step $t \in \{1, ..., T\}$. To begin with, we denote the initial state of forward process as z_0. With the constructed forward process q, we can sample $\mathbf{z}_t$ at any arbitrary noise level conditioned on $\mathbf{z}_0$, given the following formulation:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}\left(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{4}$$

where $\alpha_t =: 1 - \beta_t$ and $\bar{\alpha}_t =: \Pi_{s=1}^{t}\alpha_s$. Thus, $\mathbf{z}_t$ can be directly sampled with the reparameterization trick [33] within the forward process.

3.2.3 Reverse Process. The reverse process aims to denoise the currupted $\mathbf{z}_t$ and reconstruct the original interaction vector $\mathbf{z}_0$, which is parametrized as:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0) = \mathcal{N}\left(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)\right), \tag{5}$$

where the mean and variance are Gaussian parameters conditioned on the time step t, and can be learned through a neural network parameterized by θ.

At each time step t, the neural network can be interpreted as an equally weighted sequence of denoising autoencoders $\epsilon_\theta(\mathbf{z}_t, t); t = 1...T$, which are trained to predict a denoised variant of their input $\mathbf{z}_t$. In practice, training equivalently consists of minimizing the variational upper bound on the negative log likelihood as:

$$\mathbb{E}\left[-\log p_\theta(\mathbf{z}_0)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{z}_{0:T})}{q(\mathbf{z}_{1:T}|\mathbf{z}_0)}\right] =: \mathcal{L}_{vlb}\,, \tag{6}$$

where $\mathcal{L}_{vlb}$ can be rewritten in terms of reconstructing log-likelihood and KL divergences during each time step t, which nearly leads to $\mathcal{L}_{vlb}$ being tractable KL divergences between Gaussian distributions. Following the previous work by Ho et al. [13], we simply optimize the noise prediction loss as follows:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{z}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right], \tag{7}$$

where ϵ is sampled from a standard Gaussian for adding the noise and ϵ_θ is an approximator to predict ϵ. In this way, θ is optimized to iteratively recover $\mathbf{z}_{t-1}$ from $\mathbf{z}_t$. Finally, the overall objective of the diffusion model can be formulated as follows:

$$\begin{aligned}\mathcal{L}_t(\psi,\phi,\theta) = &- D_{KL}\left(q_\phi(\mathbf{z}|x)\|p_\psi(\mathbf{z})\right) + \mathbb{E}_{q_\phi(\mathbf{z}|x)}\left[\log p_\psi(x|\mathbf{z})\right] \\ &+ \lambda\mathbb{E}_{t,\mathbf{z}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right],\end{aligned} \tag{8}$$

where λ is a hyperparameter that controls the scale of two loss terms.

3.3 Diffusion Inference

At the inference phase, $\mathbf{z}_0$ will be obtained through encoder $\mathcal{E}$ without variance, such that $\mathbf{z}_0 = \mu_\phi(x)$. Then the low-dimensional vector z_0 will be corrupted with the forward process and reconstructed with the reverse process, and we obtain the reconstructed $\hat{z}_0$. Finally, by feeding the reconstructed $\hat{\mathbf{z}}_0$ into the decoder $\mathcal{D}$, we will obtain the augmented user-item interaction matrix $\hat{x}$.

Discussion. Different from the image diffusion models which are conducted in the pixel level, our proposed diffusion training is conducted in the latent space of user interactions, thus being suitable for recommendation. With the augmented interactions $\hat{x}$ via diffusion, the next problem becomes effectively utilizing the augmented data to benefit the downstream sequential recommendation. To tackle the problem, we then propose the curriculum scheduler.

3.4 Curriculum Scheduler

Considering that the augmented user-item interactions can play important roles in both the sequential modeling and the parameter optimization, we propose the two-level curriculum augmentation scheduler to effectively utilize the generated samples. To alleviate the impact of the potential noises hidden in the generated samples, we design an easy-to-hard curriculum training strategy.

3.4.1 Level-I: Interaction Augmentation. Primarily, the diffusion-generated samples can enrich the historical sequences, which contributes to reveal users' diversified and undiscovered intents. For the u^{th} user's historical sequence $S_u = [i_1, i_2, \cdots, i_N]$ with length N, we use the diffusion-generated matrix $\hat{x}$ to augment the user's historical sequence into K (typically 10) new historical sequences. Specifically, we first fetch all the generated behaviors of the u^{th} user $\hat{x}_u$, and then randomly replace items in S_u with sampled items from $\hat{x}_u$ at a certain $ratio$. Finally, we reorder the replaced sequence randomly to obtain the resulting augmented sequence. This augmented process is conducted K times so that we can obtain K augmented sequences $\{S'_{u,r}\}_{r=1}^K$, where each $S'_{u,r}$ is a sequence with length N, i.e., $S'_{u,r} = [i'_{1,r}, i'_{2,r}, \cdots, i'_{N,r}]$.

3.4.2 Level-II: Objective Augmentation. The augmented sequences can then be used to train the sequential recommendation model, where it is possible to choose any sequence encoder for recommendation as the backbone. Take SASRec [15] as an example, the original training objective will be the next prediction loss:

$$\mathcal{L}_{origin} = \sum_{j=1}^{N} log(p(i_{j+1}|i_1, \cdots, i_j)), \tag{9}$$

where the strategy is using previous behaviors to predict future behaviors. With the diffusion-augmented sequences, the augmented objective can be written as follows:

$$\mathcal{L}_{aug} = \mathcal{L}_{origin} + \sum_{r=1}^{K}\sum_{j=1}^{N} log(p(i'_{j+1,r}|i'_{1,r}, \cdots, i'_{j,r})), \tag{10}$$

where we additionally add the losses of next behavior prediction based on the augmented sequences to enrich the original objective.

3.4.3 Curriculum Scheduling. The augmented sequences $\{S'_{u,r}\}_{r=1}^K$ may contain noisy samples which fail to reflect the user's true intentions and may even do harm to capturing the user's intentions. We therefore design a curriculum scheduling strategy, whose key function is to denoise [42] with the easy-to-hard paradigm. Specifically, we use the similarity between the augmented sequence and the original sequence to evaluate the difficulty of each augmented sequence as follows,

$$\mathrm{Sim}_{u,r} = cos(S_u, S'_{u,r}), \tag{11}$$

where $cos(\cdot,\cdot)$ indicates the cosine similarity. More specifically, we first obtain all the item embeddings in sequences S_u and $S'_{u,r}$, and the cosine similarity $cos(S_u, S'_{u,r})$ is then calculated using the average values of embeddings belonging to each of the two sequences, which guarantees the insensitivity to the inner order of a sequence. Higher similarity implies lower difficulty. We first employ sequences with lower difficulties to optimize the model, gradually selecting those with higher difficulties for optimization.

4 EXPERIMENT

4.1 Experimental Setup

4.1.1 Datasets. We conduct extensive experiments with the proposed approach on several recommendation datasets. **Movielens-1M**: the widely used user rating dataset collected from the movielens website, containing 1,000,000 ratings from 6,000 users on 4,000 movies. **Amazon-Beauty**: the Amazon dataset is a series of datasets containing reviews and product metadata, of which Amazon-Beauty is a subset that covers rich user interactions with beauty products. **Steam**: the Steam dataset is collected from an online video game distribution platform, which encompasses extensive information about users' gaming activities, such as play hours, price, category,

Algorithm 1 The curriculum scheduling strategy for training sequential recommenders.

input: user set $\mathcal{U}$, historical sequences $\mathcal{S}$, diffusion generated interactions $\hat{x}$
output: the learned sequential recommender f_τ
function InteractAugment($S_u, \hat{x}_u, K$, ratio)
  n = ratio $* |S_u|$
  for $r = 1, ..., K$ **do**
    $S'_{u,r} = S_u$
    Replace $S'_{u,r}$ with randomly sampled n actions from $\hat{x}_u$
    Reorder $S'_{u,r}$
  return $\{S'_{u,r}\}_{r=1}^{K}$
function Sim(S_x, S_y)
  $\mathbf{x} = mean(\text{Embed}(S_x))$, $\mathbf{y} = mean(\text{Embed}(S_y))$
  $Similarity = cos(\mathbf{x}, \mathbf{y})$
  return $Similarity$
for each user u^{th} in $\mathcal{U}$ **do**
  $\{S'_{u,r}\}_{r=1}^{K} \leftarrow$ InteractAugment($S_u, \hat{x}_u, K$, ratio)
  $\{\text{Sim}_{u,r}\}_{r=1}^{K} \leftarrow$ Sim($S_u, S'_{u,r}$)
repeat
  $S \leftarrow \arg\min_{\text{Sim}_r} \{S'_r\}, \{S'_r\}.pop(S)$
  $f_\tau \leftarrow S$
  Take gradient descent step on $\nabla \mathcal{L}_{aug}$ in Eq.(10)
until *converged*

media rating, and developer details [15]. **Yelp**: the Yelp dataset is a representative business dataset containing user reviews of different restaurants, bars, etc.

The detailed descriptions and statistics of the datasets are shown in Table 1. Typically, the reviews or ratings are treated as implicit feedback, representing a user-item interaction, and organized chronologically by their associated timestamps. Additionally, users with fewer than five related actions are removed.

Table 1: Statistics of datasets. Avg. Length denotes the average length of sequences.

Dataset	# Users	# Items	# Actions	Avg. Length	Sparsity
Movielens-1M	6,041	3,707	1,000,209	163.5	95.53%
Beauty	40,226	54,542	353,962	8.8	99.98%
Steam	281,428	13,044	3,485,022	12.4	99.90%
Yelp	30,431	20,033	1,301,869	12.1	99.79%

4.1.2 Evaluation Settings. The performance evaluation is conducted by the widely used *leave-one-out* strategy [35, 39, 47], where the most recent interaction is kept as the test data, the penultimate interaction is for validation, and all earlier interactions are for training. We employ top-k Hit Ratio (HR@k), top-k Normalized Discounted Cumulative Gain (NDCG@k) and Mean Reciprocal Rank (MRR) to evaluate the recommendation performance, which are widely used as common practice [15, 26, 44]. We report results on HR@{5, 10, 20} and NDCG@{5, 10}. To ensure consistent evaluation results, we rank all candidate items for predicting the target item instead of using the biased sampling, especially in cases where the number of negative items is small [18].

4.1.3 Comparison Baselines. We compare Diff4Rec against a series of representative baselines, including conventional recommendation models, state-of-the-art sequential models and recently proposed novel approaches.

BPR [29] is a matrix factorization variant of the classic Bayesian personalized ranking algorithm. **NCF** [10] is one of the most representative collaborative filtering methods based on neural networks. **GRU4Rec** [12] firstly employs the GRU network with ranking based loss for sequential recommendation. **Caser** [36] employs CNN in both horizontal and vertical way to model the user's subsequence behaviors. **SASRec** [15] utilizes self-attention [37] to exploit the long-term mutual influence between historical interactions. **BERT4Rec** [35] further designs a bidirectional Transformer with cloze task for sequential recommendation. **S^3Rec** [47] is a self-supervised learning method which devises four pretext tasks for context-aware recommendation and then finetunes on the next-item recommendation task. **STOSA** [6] is a recently proposed uncertainty recommendation model that employs a Wasserstein self-attention to consider collaborative transitivity in sequential recommendation. **ContraRec** [39] is a contrastive learning based method which leverages two signals named *context-target contrast* and *context-context contrast* for sequential recommendation.

4.2 Overall Performance

We first present the overall comparison between Diff4Rec and baseline models in Table 2, from which we can summarize the observations as follows.

Diff4Rec shows consistent superior performance over all the baseline models across four datasets in terms of all metrics. In particular, compared with the different streams of baselines, Diff4Rec achieves impressive improvements on each dataset. The improvement is especially impressive on the sparse Beauty and Steam datasets, where the relative improvement over the strongest baselines is in general over 20%. Such improvements suggest that 1) Diff4Rec is capable of modeling intricate patterns of user-item interactions by progressively learning from every denoising transition step. 2) The diffusion model learned in the latent space effectively fits various latent intentions of users. 3) The curriculum scheduling strategy effectively augment the sequential behaviors with well selected data and encourages the model to optimize within a larger scope to yield better recommendation.

4.3 Ablation Studies

4.3.1 Comparison on Base Sequential Models. Since Diff4Rec serves as a general framework that is capable of being applied on different sequential recommenders, we evaluate it by integrating three representative techniques for sequential interaction modeling, including RNN (GRU4Rec [12]), CNN (Caser [36]), and Transformer (SASRec [15]), as the base sequence encoder. We conduct the comparison on three datasets and the experimental results are shown in Table 3. The results demonstrate the steady and superior performance of Diff4Rec upon multiple base sequential models, where Diff4Rec achieve prominent improvements on all the datasets. The consistent superiority reflects the effectiveness and flexibility of our proposed framework.

Table 2: Overall performance comparisons with baseline models on four datasets (the best results are in BOLD, second in underline). Improve. represents the relative improvements of Diff4Rec over the best baseline results.

Dataset	Metric	BPR	NCF	GRU4Rec	Casar	SASRec	BERT4Rec	S^3Rec	STOSA	ContraRec	Diff4Rec	Improv.
ML-1M	HR@5	0.0155	0.0147	0.0792	0.0813	0.1047	0.1128	0.1154	0.0624	0.1269	**0.1401**	10.42%
	HR@10	0.0327	0.0309	0.1573	0.1556	0.1775	0.1840	0.1921	0.1347	0.1980	**0.2238**	13.02%
	HR@20	0.0649	0.0651	0.2374	0.2528	0.2807	0.3083	0.3199	0.2580	0.3423	**0.3830**	11.90%
	NDCG@5	0.0086	0.0072	0.0458	0.0512	0.0598	0.0649	0.0640	0.0297	0.0725	**0.0812**	12.03%
	NDCG@10	0.0140	0.0124	0.0667	0.0740	0.0824	0.1041	0.1077	0.0684	0.0987	**0.1222**	13.45%
	MRR	0.0100	0.0090	0.0479	0.0673	0.0819	0.0984	0.1062	0.0681	0.0815	**0.1202**	13.18%
Beauty	HR@5	0.0121	0.0146	0.0152	0.0158	0.0325	0.0247	0.0307	0.0325	0.0385	**0.0456**	18.35%
	HR@10	0.0277	0.0277	0.0268	0.0263	0.0633	0.0396	0.0574	0.0684	0.0753	**0.0871**	15.62%
	HR@20	0.0496	0.0486	0.0413	0.0427	0.0971	0.0555	0.0933	0.0910	0.1099	**0.1347**	22.54%
	NDCG@5	0.0063	0.0078	0.0094	0.0109	0.0204	0.0125	0.0203	0.0248	0.0221	**0.0301**	21.50%
	NDCG@10	0.0119	0.0121	0.0137	0.0149	0.0303	0.0160	0.0290	0.0335	0.0358	**0.0426**	18.99%
	MRR	0.0090	0.0104	0.0133	0.0113	0.0285	0.0158	0.0236	0.0292	0.0322	**0.0368**	14.42%
Steam	HR@5	0.0103	0.0111	0.0340	0.0396	0.0464	0.0478	0.0469	0.0481	0.0510	**0.0575**	12.67%
	HR@10	0.0239	0.0258	0.0561	0.0680	0.0880	0.0851	0.0891	0.0844	0.1003	**0.1132**	12.90%
	HR@20	0.0427	0.0465	0.0905	0.1011	0.1316	0.1240	0.1350	0.1426	0.1451	**0.1608**	10.79%
	NDCG@5	0.0057	0.0066	0.0126	0.0227	0.0224	0.0284	0.0288	0.0272	0.0261	**0.0321**	11.32%
	NDCG@10	0.0101	0.0102	0.0268	0.0315	0.0402	0.0429	0.0424	0.0421	0.0489	**0.0541**	10.70%
	MRR	0.0092	0.0089	0.0213	0.0296	0.0350	0.0390	0.0387	0.0405	0.0434	**0.0479**	10.37%
Yelp	HR@5	0.0126	0.0137	0.0142	0.0139	0.0153	0.0158	0.0174	0.0182	0.0188	**0.0198**	5.58%
	HR@10	0.0244	0.0256	0.0252	0.0247	0.0286	0.0291	0.0300	0.0395	0.0303	**0.0411**	4.06%
	HR@20	0.0470	0.0449	0.0453	0.0451	0.0481	0.0488	0.0479	0.0523	0.0573	**0.0595**	3.90%
	NDCG@5	0.0078	0.0074	0.0087	0.0083	0.0089	0.0094	0.0102	0.0114	0.0103	**0.0125**	9.25%
	NDCG@10	0.0118	0.0109	0.0124	0.0122	0.0134	0.0146	0.0153	0.0157	0.0160	**0.0171**	6.83%
	MRR	0.0089	0.0085	0.0101	0.0103	0.0119	0.0133	0.0141	0.0146	0.0144	**0.0157**	7.59%

Table 3: Performance comparison of Diff4Rec with different base sequence encoders. NG@k is short for NDCG@k.

Method	Beauty		Steam		Yelp	
	HR@10	NG@10	HR@10	NG@10	HR@10	NG@10
GRU4Rec	0.0268	0.0137	0.0561	0.0268	0.0252	0.0124
+Diff4Rec	0.0372	0.0190	0.0691	0.0342	0.0336	0.0161
Caser	0.0263	0.0149	0.0680	0.0315	0.0247	0.0122
+Diff4Rec	0.0375	0.0214	0.0838	0.0443	0.0351	0.0161
SASRec	0.0633	0.0303	0.0880	0.0402	0.0286	0.0134
+Diff4Rec	**0.0871**	**0.0426**	**0.1132**	**0.0541**	**0.0411**	**0.0171**

Table 4: Ablation study on the curriculum scheduling strategies of Diff4Rec. NG@k is short for NDCG@k.

Method	Beauty		Steam		Yelp	
	HR@10	NG@10	HR@10	NG@10	HR@10	NG@10
Base	0.0633	0.0303	0.0880	0.0402	0.0286	0.0134
Diff4Rec$_{w/o\ \text{cur}}$	0.0850	0.0417	0.1108	0.0524	0.0410	0.0165
Diff4Rec$_{w/\text{cur}}$	**0.0871**	**0.0426**	**0.1132**	**0.0541**	**0.0411**	**0.0171**

4.3.2 Curriculum Scheduling Strategies. To validate the effectiveness of our proposed curriculum scheduler, we conduct ablation studies on the curriculum scheduling strategies of Diff4Rec. Consistently, we adopt SASRec [15] as the base sequential recommender, and then trial two types of implementation of Diff4Rec:

- Diff4Rec$_{w/o\ \text{cur}}$ utilizes the *interaction augmentation* and *objective augmentation* to learn the sequential recommender while not performing curriculum scheduling selection on the augmented samples.
- Diff4Rec$_{w/\text{cur}}$ utilizes both curriculum scheduling and augmentation, which is essentially our presented Diff4Rec.

We compare the performance of these implementations on three datasets and the experimental results are shown in Table 4, from which we can conclude 1) Diff4Rec$_{w/o\ \text{cur}}$ performs much better over base model in all cases, which indicates the significance of the data generation and two-level augmentation in Diff4Rec. 2) Diff4Rec$_{w/\text{cur}}$ achieves best results in all cases, showing the efficacy and necessity of the designed curriculum scheduler. The results reflect that the Diff4Rec properly selects out augmented samples which are positive to augment sequence features, and the *interaction augmentation* and *objective augmentation* performed by the curriculum scheduler effectively enrich sparse interactions and result in better recommendation.

5 CONCLUSION

In this paper, we study the problem of utilizing diffusion models to conduct data augmentation for sequential recommendation. We bravely propose the Diff4Rec framework, which consists of a diffusion model to learn user-item interactions in the latent space and a curriculum-guided augmentation strategy to schedule the diffusion-generated samples. Experiments demonstrate that our proposed Diff4Rec can effectively bring significant improvements to the sequential recommendation. Utilizing diffusion models to enhance the sequential recommendation could be a promising future research direction.

ACKNOWLEDGMENTS

This work is supported in part by National Key Research and Development Program of China No.2022ZD0115903, NSFC No. 62222209, 62250008, 62102222, BNRist under Grant No. BNR2023RC01003, BNR2023TD03006, and Beijing Key Lab of Networked Multimedia.

REFERENCES

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[2] Shuqing Bian, Wayne Xin Zhao, Kun Zhou, Jing Cai, Yancheng He, Cunxiang Yin, and Ji-Rong Wen. 2021. Contrastive curriculum learning for sequential user behavior modeling via data augmentation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3737–3746.

[3] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2021. WaveGrad: Estimating Gradients for Waveform Generation. In *International Conference on Learning Representations*.

[4] Yudong Chen, Xin Wang, Miao Fan, Jizhou Huang, Shengwen Yang, and Wenwu Zhu. 2021. Curriculum meta-learning for next POI recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2692–2702.

[5] Zhengxiao Du, Xiaowei Wang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Sequential scenario-specific meta learner for online recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2895–2904.

[6] Ziwei Fan, Zhiwei Liu, Yu Wang, Alice Wang, Zahra Nazari, Lei Zheng, Hao Peng, and Philip S Yu. 2022. Sequential recommendation via stochastic self-attention. In *Proceedings of the ACM Web Conference 2022*. 2036–2047.

[7] Ziwei Fan, Zhiwei Liu, Jiawei Zhang, Yun Xiong, Lei Zheng, and Philip S Yu. 2021. Continuous-time sequential recommendation with temporal graph collaborative transformer. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 433–442.

[8] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. 2020. Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations. *ACM Transactions on Information Systems (TOIS)* 39, 1 (2020), 1–42.

[9] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 191–200.

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[11] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 843–852.

[12] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, Vol. 33. 6840–6851.

[14] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. 2022. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems*, Vol. 35. 24240–24253.

[15] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[16] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[17] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations*.

[18] Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1748–1757.

[19] Hui Li, Ye Liu, Nikos Mamoulis, and David S Rosenblum. 2019. Translation-based sequential recommendation for complex users on sparse data. *IEEE Transactions on Knowledge and Data Engineering* 32, 8 (2019), 1639–1651.

[20] Haoyang Li, Xin Wang, Ziwei Zhang, Jianxin Ma, Peng Cui, and Wenwu Zhu. 2021. Intention-aware sequential recommendation with structured intent transition. *IEEE Transactions on Knowledge and Data Engineering* 34, 11 (2021), 5403–5414.

[21] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1419–1428.

[22] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*. 689–698.

[23] Shitong Luo and Wei Hu. 2021. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2837–2845.

[24] Shitong Luo, Chence Shi, Minkai Xu, and Jian Tang. 2021. Predicting molecular conformation via dynamic graph score matching. In *Advances in Neural Information Processing Systems*, Vol. 34. 19784–19795.

[25] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. In *Advances in Neural Information Processing Systems*, Vol. 32.

[26] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled self-supervision in sequential recommenders. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 483–491.

[27] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*. PMLR, 8599–8608.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[29] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.

[30] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[32] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–10.

[33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

[34] Yang Song and Stefano Ermon. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

[35] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 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.

[36] Jiaxi Tang and Ke Wang. 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.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30.

[38] Joojo Walker, Ting Zhong, Fengli Zhang, Qiang Gao, and Fan Zhou. 2022. Recommendation via Collaborative Diffusion Generative Model. In *Knowledge Science, Engineering and Management: 15th International Conference, KSEM 2022, Singapore, August 6–8, 2022, Proceedings, Part III*. Springer, 593–605.

[39] Chenyang Wang, Weizhi Ma, Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. Sequential recommendation with multiple contrast signals. *ACM Transactions on Information Systems* 41, 1 (2023), 1–27.

[40] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2015. Learning hierarchical representation model for nextbasket recommendation. In *Proceedings of the 38th International ACM SIGIR conference on Research and Development in Information Retrieval*. 403–412.

[41] Wenjie Wang, Yiyan Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. 2023. Diffusion Recommender Model. *arXiv preprint arXiv:2304.04971* (2023).

[42] Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 4555–4576.

[43] Zhenlei Wang, Jingsen Zhang, Hongteng Xu, Xu Chen, Yongfeng Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Counterfactual data-augmented sequential recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 347–356.

[44] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 1259–1273.

[45] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. 2022. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*.

[46] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 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.

[47] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1893–1902.