Conditional Denoising Diffusion for Sequential Recommendation

Yu Wang[1(✉)], Zhiwei Liu[2], Liangwei Yang[1], and Philip S. Yu[1]

[1] University of Illinois Chicago, Chicago, USA
{ywang617,lyang84,psyu}@uic.edu
[2] Salesforce AI Research, Palo Alto, USA
zhiweiliu@salesforce.com

Abstract. Contemporary attention-based sequential recommendations often encounter the oversmoothing problem, which generates indistinguishable representations. Although contrastive learning addresses this problem to a degree by actively pushing items apart, we still identify a new *ranking plateau* issue. This issue manifests as the ranking scores of top retrieved items being too similar, making it challenging for the model to distinguish the most preferred items from such candidates. This leads to a decline in performance, particularly in top-1 metrics. In response to these issues, we present a conditional denoising diffusion model that includes a stepwise diffuser, a sequence encoder, and a cross-attentive conditional denoising decoder. This approach streamlines the optimization and generation process by dividing it into simpler, more tractable sub-steps in a conditional autoregressive manner. Furthermore, we introduce a novel optimization scheme that incorporates both cross-divergence loss and contrastive loss. This new training scheme enables the model to generate high-quality sequence/item representations while preventing representation collapse. We conduct comprehensive experiments on four benchmark datasets, and the superior performance achieved by our model attests to its efficacy. We open-source our code at https://github.com/YuWang-1024/CDDRec.

Keywords: Sequential Recommendation · Diffusion Models · Generative Models

1 Introduction

Sequential Recommendation (SR) [10,13,24–26,28] has been intensively investigated because of its scalability and efficacy in capturing user temporal trends from histories. Recent research in SR focuses on attention-based methods for their promising results. Early attempts e.g., SASRec [10] and Bert4Rec [22] utilize the attention-based transformer structure. However, the attention mechanism tends to lead to a condition known as oversmoothing [5,6], which results in generating indistinguishable representations. Current methods predominantly

D.-N. Yang et al. (Eds.): PAKDD 2024, LNAI 14649, pp. 156–169, 2024.
https://doi.org/10.1007/978-981-97-2262-4_13

address this complication from the item representation perspective, utilizing contrastive learning. These methods effectively counter the collapse of item representation learning [17,26,27,29]. The DuoRec model [17] proposes to use sequences with the same predicted item as augmented views and implements the noise contrastive estimation objective for regularization. ContrastVAE [26] incorporates variational augmentation and the contrastELBO objective into the attention-based variational autoencoder for SR.

Despite the success of the above methods, we still observe a phenomenon we term the *ranking plateau*, characterized by indistinguishable ranking scores, even when the quality of item representation is commendable. Specifically, as shown in Sect. 4.1, the ranking scores of the top-40 retrieved items are too similar to allow the recommender to differentiate the best candidates among them. These oversmoothed ranking scores result in performance degradation, especially in top-1 metrics. For instance, the score assigned by DuoRec to the second-best item is only 1% lower than that of the top item. (We will discuss such phenomenon in detail in Sect. 4.1.) This suggests that the under-performance might not be solely due to item representation degeneration, but also to the complicated reasons beyond the user-item engagement. During our experiments, we observe significant shifts in some user intents from their historical records. For example, if a user regularly purchases *collection kits* but suddenly transitions to buying *printing-related* items, the representation of such sequences can be easily skewed by the user's past behaviors. This results in continuously recommending *collection-related* items, as the sequence representations are essentially a weighted sum of past item representations, regardless of the quality of candidate item representations.

Intuitively, if such dynamic intent transitions cannot be captured during the one-step dot-product, one might question whether it would be feasible to divide the transition process into easier and more tractable multi-steps such that the model could potentially correlate the intermediate transition steps and produce high-fidelity results progressively. For these desiderata, we turn to diffusion models [3,8,18] for solutions, as they break the higher-order complicated transitions into feasible sub-steps by removing certain noise stepwisely. Generally, the diffuser of diffusion models gradually adds a certain scale of Gaussian Noise to the data in the forward diffusion process, and the denoiser reconstructs such intermediate states by learning to remove the added noise in the reverse denoising process. In this way, the denoiser is able to learn fine-grained intermediate transitions from these multi-step generations.

However, it is rather challenging to incorporate such a learning paradigm into SR. One primary reason is that traditional diffusion models are designed for continuous spaces like image generation, where input features are fixed and contain substantial information. They are optimized by reconstructing **original** images. In contrast, SR involves item input information that is randomly initialized based on item IDs and dynamically optimized. Original reconstruction objective in discrete spaces could be adversely affected by representation collapse that all embeddings collapse to a trivial solution [3,4,11]. Furthermore, the SR is a retrieval task, which aims to generate user preferences reflecting the **next**

item engagement. Merely reconstructing original item representations within a sequence (Gaussian Noise vector from the beginning) does not contribute to ranking performance but exacerbates the collapse issues of diffusion models.

To address these challenges, we propose **C**onditional **D**enoising **D**iffusion Models for Sequential **Rec**ommendation (CDDRec), including a stepwise diffuser, sequence encoder, cross-attentive conditional denoising decoder, and cross-divergence objective. The stepwise diffuser introduces noise into target item representations to construct corrupted targets, simulating the small stepwise noise in the sequences. The sequence encoder learns sequence representations from historical interactions, used as the conditioned information for a stepwise generation of next-engagement preference. The conditional denoising decoder aims to generate high-quality next-engagement representation by removing the noise of historical sequence representations step-by-step. To enhance the denoising decoder's awareness of each denoising step, we adopt the cross-attention mechanism with the denoising step as input. Additionally, We introduce a cross-divergence loss, enabling the model to construct high-fidelity sequence/item representations while being attuned to next-engagement preferences and preventing learning collapse. Furthermore, we leverage the In-view and Cross-view contrastive optimization to prevent item representation degeneration. Our contribution can be summarized as follows:

- To the best of our knowledge, we are the first to propose the novel conditional denoising diffusion models for sequential recommendation CDDRec in the conditional autoregressive generation paradigm.
- We first observe the *ranking plateau* issue and propose the multi-step next-engagement generation to address this issue.
- We introduce cross-divergence to equip the CDDRec with ranking capability.
- We conduct comprehensive experiments on the SR dataset, the substantial improvement across all metrics in four datasets indicates the effectiveness of CDDRec. We also conduct ablation studies to examine each key design's effectiveness further.

2 Related Work

Denoising Diffusion Probabilistic Models (DDPMs) have shown great success in continuous spaces, such as image generation [8,9,15,18]. Recently, several attempts have been made to apply DDPMs to discrete tasks, such as text generation. SUNDAE [21] is one of pioneers that use DDPMs for text generation. They introduce a step-unrolled denoising autoencoder that reconstructs corrupted sequences in a non-autoregressive manner. Diffusion-LM [11] gradually reconstructs word vectors from Gaussian noise guided by attribute classifiers and introduces a rounding process that maps continuous word embeddings to discrete words. DiffSeq [4] introduces a forward process with partial noise that uses the question of a dialog as the uncorrupted part and the answer of the dialog as the corrupted part and adds partial noise to the answer part during the forward pass. The backward pass reconstructs the answer in a non-autoregressive way.

There are also several concurrent attempts to introduce DDPMs for recommender systems. [23] introduces noise and reconstructs information for user interactions, and introduces L-DiffRec resembling latent diffusion, and T-DiffRec to encode temporal information to reweight user interactions respectively. [12] adds Gaussian noise on target items and reconstructs them through an approximator, which inputs the corrupted target item representations and historical interactions. [2] introduces the partial noise only on the target items that resemble DiffSeq and reconstructs them in a non-autoregressive way. Unlike the above methods, we introduce conditional generation in an autoregressive manner, which equips the model with the ability to generate high-fidelity sequence/item representations without too many generation steps.

3 Methodology

In this section, we present the methodology of CDDRec and illustrate it in Fig. 1. CDDRec consists of 1) stepwise diffuser that gradually corrupts target item embeddings via adding Gaussian noise; 2) sequence encoder that learns historical sequence representation, serving as conditional information for stepwise preference generation; 3) cross-attentive conditional denoising decoder that learns the stepwise user preference transition from conditional historical sequence representation to next target preference via stepwise removing noise from conditional sequence embeddings; 4) cross-divergence objective that enables the model with ranking capability while preventing model from collapsing.

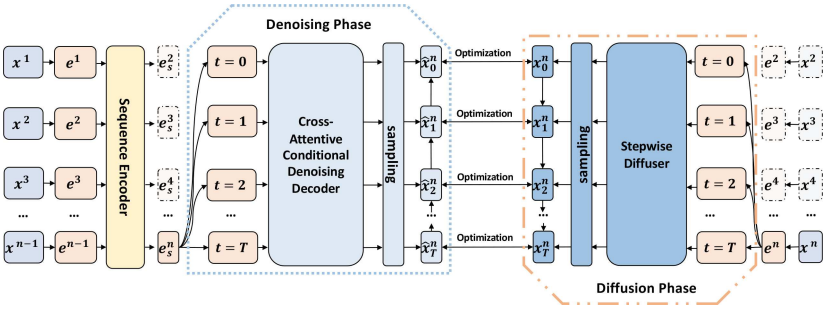


Fig. 1. Framework of CDDRec. Orange dots from top to bottom indicate the diffusion phase that gradually adds Gaussian noise to target item embeddings $\mathbf{x_t^n}$, while blue dots from the bottom up illustrate the reverse denoising phase that stepwisely removes noise from estimated user preference $\hat{\mathbf{x}}_\mathbf{t}^\mathbf{n}$ at step t. (Color figure online)

3.1 Stepwise Diffuser

As shown in Fig. 1 with the orange dot box from top to bottom, the stepwise diffuser is designed to incrementally introduce Gaussian noise to target item embeddings. This process creates the corrupted target for each step, thereby

facilitating the denoise learning of the denoiser. Given the predefined noise scale added at diffusion step t: β_t and the corresponding diffusion transition distribution $q(\mathbf{x_t^n}|\mathbf{x_{t-1}^n}) \sim \mathcal{N}(\mathbf{x_t^n}; \sqrt{1-\beta_t}\mathbf{x_{t-1}^n}, \beta_t\mathbf{I})$, the distribution of the current diffusion step has the analytical form conditional on the first diffusion step:

$$q(\mathbf{x_t^n}|\mathbf{x_0^n}) = \mathcal{N}(\mathbf{x_t^n}; \sqrt{\bar{\alpha}_t}\mathbf{x_0^n}, (1-\bar{\alpha}_t)\mathbf{I}), \quad \alpha_t = 1-\beta_t, \quad \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i, \tag{1}$$

where $\mathbf{x_0^n} = \mathbf{e^n}$ is target item representation as the initialization of the diffusion phase. Thus, we can sample the corrupted target $\mathbf{x_t^n}$ at any step t using $\mathbf{x_0^n}$:

$$\mathbf{x_t^n} = \sqrt{\bar{\alpha}_t}\mathbf{x_0^n} + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{2}$$

3.2 Sequence Encoder

Previous methods predominantly concentrate on denoising from a randomly initialized Gaussian noise and generating sentences non-autoregressively [3,4,11]. However, predicting the next item based on historical interaction records necessitates a conditional autoregressive generation in the SR. Consequently, in this paper, we utilize SASRec as our sequence encoder to learn hidden representations of historical interactions $\mathbf{e_s}$. These are used as the condition of the subsequent conditional denoising decoder for the multi-step preference generation.

3.3 Cross-Attentive Conditional Denoising Decoder

Given the distribution of diffusion step $q(\mathbf{x_t^n}|\mathbf{x_{t-1}^n})$, $q(\mathbf{x_t^n}|\mathbf{x_0^n})$, and $q(\mathbf{x_{t-1}^n}|\mathbf{x_0^n})$ as described in Sect. 3.1, we can compute the analytical form of posterior distribution using Bayes' rule: $q(\mathbf{x_{t-1}^n}|\mathbf{x_t^n}, \mathbf{x_0^n})$, which is the reverse denoising distribution, with $\hat{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ as the closed form of variance. We approximate such reverse denoising step with distribution $p_\theta(\hat{\mathbf{x}}_{\mathbf{t-1}}^{\mathbf{n}}|\hat{\mathbf{x}}_{\mathbf{t}}^{\mathbf{n}}) \sim \mathcal{N}(\mu_\theta(\hat{\mathbf{x}}_{\mathbf{t}}^{\mathbf{n}}, t), \hat{\beta}_t\mathbf{I})$, parameterized by learnable parameter θ that learns the denoised representation $\hat{\mathbf{x}}_{\mathbf{t-1}}^{\mathbf{n}}$ at step $t-1$ conditional on the previous denoising step $\hat{\mathbf{x}}_{\mathbf{t}}^{\mathbf{n}}$. For SR tasks, the objective is to predict subsequent items based on historical interactions in an autoregressive manner. Therefore, rather than generating sequence representations from uncontrollable randomly initialized Gaussian noise, we integrate the denoiser within the conditional generation framework with a conditional denoising decoder. Consequently, as shown in Fig. 1 blue dot box from bottom up, we condition the reverse denoising phase on the preceding sequence representations, expressed as $p_\theta(\hat{\mathbf{x}}_\mathbf{t}|\mathbf{e_s}, t) \sim \mathcal{N}(\mu_\theta(\mathbf{e_s}, t), \hat{\beta}_{t+1}\mathbf{I})$, where $\mathbf{e_s}$ denotes the encoded historical interactions using the sequence encoder. Given such probabilistic modeling, we will discuss the corresponding model design to learn the denoised mean $\mu_\theta(\mathbf{e_s}, t)$.

In contrast to earlier methods [19] that only maintain the final position's representation as the sequence representation, we strive to preserve as much information as possible due to the sparse nature of SR. Hence, we select a cross-attention architecture as the denoising decoder instantiation, which is capable

of taking the entire sequence representation and corresponding step indicator as input. Formally, given a sequence embedding $\mathbf{e}_\mathbf{s}^{\mathbf{1:n}}$ and the corresponding denoising step t, the conditional denoising decoder is designed to predict the denoised mean of corrupted target item embedding at the corresponding diffusion step. Initially, we acquire a learnable embedding $\mathbf{e_t}$ for the indicator t from a step lookup embedding table and expand it to the dimension of $\mathcal{R}^{(n-1)\times d}$, ensuring that every previously hidden embedding is conscious of the same denoising step. We define the cross-attention (CA) as follows:

$$\mu_\theta^{\mathbf{1:n}}(\mathbf{e}_\mathbf{s}^{\mathbf{1:n}}, t) = CA(\mathbf{e}_\mathbf{s}^{\mathbf{1:n}}, \mathbf{e_t}) = \text{Softmax}\left(\frac{(\mathbf{e_t}\mathbf{W}^Q)(\mathbf{e}_\mathbf{s}^{\mathbf{1:n}}\mathbf{W}^K)^\top}{\sqrt{d}}\right)(\mathbf{e}_\mathbf{s}^{\mathbf{1:n}}\mathbf{W}^V). \tag{3}$$

Given the predicted denoised mean and the precomputed posterior variance, we can sample the generated user preference at step t:

$$\hat{\mathbf{x}}_\mathbf{t}^\mathbf{n} = \mu_\theta^\mathbf{n} + \hat{\beta}_{t+1}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{4}$$

3.4 Optimization

Traditional DDPM is designed to reconstruct an image by removing the Gaussian noise added to it. Consequently, the objective is to learn the denoising function $p_\theta(x_{t-1}|x_t)$ for the corresponding step, minimizing the KL divergence $D_{KL}[q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)]$ at each step. Such an objective maximizes the similarity between the predicted and corrupted input data. However, since all item embeddings are randomly initialized and optimized dynamically in SR, the model may learn trivial item representations, where every pair of item embeddings is highly similar, resulting in high-ranking scores for all items. Furthermore, the SR is a retrieval task requiring the model to effectively rank items, giving higher scores to target items over non-interest items. Merely reconstructing the input sequence does not contribute to the next-engagement prediction.

To circumvent these issues, we require the KL divergence between the predicted and target item embeddings to be smaller than that between the predicted and negative item embeddings. Consequently, we introduce the cross-divergence loss using KL-divergence as a dissimilarity metric at each denoising step t:

$$\begin{aligned}\mathcal{L}_{cd}^t = &\frac{1}{N}\sum_n \log(\sigma(-D_{KL}[q(\mathbf{x_t^n}|\mathbf{x_{t+1}^n}, \mathbf{x_0^n})||p_\theta(\hat{\mathbf{x}}_\mathbf{t}^\mathbf{n}|\mathbf{e_s}, t)])) \\ &+ \log(1-\sigma(-D_{KL}[q(\mathbf{x_t'^n}|\mathbf{x_{t+1}'^n}, \mathbf{x_0'^n})||p_\theta(\hat{\mathbf{x}}_\mathbf{t}^\mathbf{n}|\mathbf{e_s}, t)]))],\end{aligned} \tag{5}$$

where $\mathbf{x_0'^n}$ is the embedding of a randomly sampled negative item that has never appeared in the user history. We sample both corrupted target item embeddings $\mathbf{x_t^n}$ and generated user preferences $\hat{\mathbf{x}}_\mathbf{t}^\mathbf{n}$ according to Eq. 2 and Eq. 4.

Contrastive Loss. To endow the model with robustness against the noisy interactions and prevent item representation from collapsing, we incorporate a simple yet effective in-view and cross-view contrastive learning using InfoNCE loss [16].

The in-view InfoNCE minimizes the distance between user preferences and target item embedding while enlarging inter-users/inter-item distance:

$$\mathcal{L}_{in}^{t} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\hat{\mathbf{x}}_t^{i\top}\mathbf{x}_t^{i}/\tau)}{\sum_j \exp(\hat{\mathbf{x}}_t^{i\top}\mathbf{x}_t^{j}/\tau) + \sum_j \mathbb{1}_{[j\neq i]}\exp(\hat{\mathbf{x}}_t^{i\top}\hat{\mathbf{x}}_t^{j}/\tau)}, \tag{6}$$

where $\hat{\mathbf{x}}_{\mathbf{t}}^{\mathbf{i}}$ is the output of conditional denoising decoder, $\mathbf{x}_{\mathbf{t}}^{\mathbf{i}}$ is the output of stepwise diffuser at diffusion step t of position i in the sequence.

The cross-view InfoNCE ensures the sequence encoder generates reasonable sequence representation. It achieves this by minimizing the distance between the same input sequence with a slight noise interpolation while pushing the in-batch sequence representation away from each other:

$$\mathcal{L}_{cross}^{t} = \frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\hat{\mathbf{x}}_t^{i\top}\tilde{\mathbf{x}}_t^{i}/\tau)}{\sum_j \exp(\hat{\mathbf{x}}_t^{i\top}\tilde{\mathbf{x}}_t^{j}/\tau) + \sum_j \mathbb{1}_{[j\neq i]}\exp(\hat{\mathbf{x}}_t^{i\top}\hat{\mathbf{x}}_t^{j}/\tau)}, \tag{7}$$

where $\tilde{\mathbf{x}}_t^i$ is the output of conditional denoising decoder at diffusion step t, position i of augmented view.

Step-Adaptive Objective. Since the noise added to target item embeddings increases as the diffusion phase progresses, more information is lost at higher diffusion steps. Intuitively, to avoid focusing too much on reconstructing non-informative noise, we rescale the loss term of each diffusion step by dividing it by the corresponding step indicator. Furthermore, unlike previous methods that randomly sample step indicators for optimization, we explicitly calculate the loss term for every diffusion step. The final optimization objective is formalized as:

$$\mathcal{L} = \sum_{t=0}^{T}\frac{1}{t+1}(\mathcal{L}_{cd}^{t} + \lambda(\mathcal{L}_{in}^{t} + \mathcal{L}_{cross}^{t})). \tag{8}$$

4 Experiments

Dataset. In this paper, we conduct experiments on four Amazon datasets [14]: *Office*, *Beauty*, *Tools and Home*, and *Toys and Games*. In line with common practice [10,22,29], for each user, we sort the interactions chronologically. We use the penultimate, last records as validation and test datasets, while all preceding records as train datasets. We report the statistics of the datasets in Table 1.

Table 1. Statistics of datasets.

Dataset	#Users	#Items	#Interactions	#Ints/item	Avg. seq. len.
Beauty	22,363	12,101	198,502	16.40	8.3
Toys	19,412	11,924	167,597	14.06	8.6
Tools	16,638	10,217	134,476	13.16	8.1
Office	4,905	2,420	53,258	22.00	10.8

Baseline Models. We compare CDDRec with these three related types of state-of-the-art (SOTA) methods: Generative Models: **SAVE** [20], **ACVAE** [28], **ContrastVAE** [26]. SVAE first introduces VAE into the SR. ACVAE introduces the concept of adversarial variational Bayes and mutual information maximization to optimize the VAE. ContrastVAE introduces the objective named ContrastELBO to maximize the mutual information among latent variables. Contrastive Models: **CL4Rec** [27], **DuoRec** [17], **CBiT** [1]. CL4Rec introduces the data augmentation strategies: mask, shuffle, and crop, and optimizes the model via InfoNCE [16] loss. DuoRec improves the performance via semantic augmentation considering sequences with the same target items as the positive views. CBiT improves Bert4Rec by introducing the additional InfoNCE objective. Encoder Models: **GRU4Rec** [7], **SASRec** [10], **FMLP** [30]. GRU4Rec first attempts the RNN for the SR, while SASRec first employs a transformer-based encoder for SR. FMLP replaces the multi-head self-attention layer of SASRec with the denoising Fourier layer.

Metrics. To evaluate the performance of our model, we employ ranking-related evaluation metrics, including Recall@N, NDCG@N, and MRR, following common practice [10,26,27].

4.1 Plateau of Ranking Prediction

As previously mentioned, traditional generative models often encounter *ranking plateau*, where the ranking scores of top-40 candidate items are too similar. This smoothness makes it difficult for models to distinguish the best from these candidates, resulting in degraded top-1 metrics. To study such a phenomenon, we conduct experiments comparing the average absolute percentage change (Avg.Change) of the top-40 ranking scores. The metric is defined as follows:

$$\text{Avg.Change} = \sum_{i=1}^{N} \frac{1}{N-1} \frac{|rank_{i+1} - rank_i|}{rank_i} \times 100, \tag{9}$$

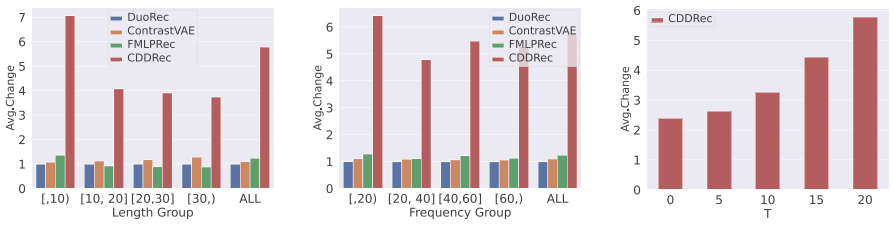
where *rank* is the ranking score calculated using dot-product between predicted and candidate item embeddings.

We utilize this metric to evaluate the descending speed of the ranking scores, which can reflect the smoothness of the ranking prediction. We report the

results in Fig. 2. In general, baseline models tend to provide more similar ranking scores among the top-40 candidates. Interestingly, we observe a positive correlation between Recall@1 and Avg.Change when comparing Avg.Change on 'ALL' sequences. Specifically, the Recall@1 is 0.012, 0.0194, 0.0224, and 0.0271 for DuoRec, ContrastVAE, FMLP, and CDDRec, and the Avg.Change is 0.99%, 1.0928%, 1.2269%, and 5.7765% respectively. The Avg.Change of CDDRec decreases with increasing sequence length, indicating that the model is less certain for longer sequences. We also evaluate the Avg.Change w.r.t. the denoising stage (inverse process w.r.t diffusion step t, i.e., denoising stage is $T-t$). One observation is that the Avg.Change increases with the denoising stage. At the beginning of the denoising stage, the model shows uncertainty in ranking predictions, but it gradually gains clarity as the denoising phase progressively removes noise from the preference predictions. This also reveals the relationship between the noise level in the generated user preference and the ranking score smoothness. Specifically, a noisier preference correlates with a smoother ranking score, which supports our intuition of adding a denoiser after the sequence encoder.

4.2 Overall Experiments

In this paper, we conduct a comprehensive comparison between CDDRec and SOTA models, reporting the numerical results in Table 2. Our model CDDRec consistently outperforms others on these four datasets, demonstrating the effectiveness of CDDRec. Specifically, in terms of Recall@1, CDDRec shows substantial improvements with gains of 20.98%, 16.67%, 17.59%, and 18.42% compared to the second-best models on Office, Beauty, Tools, and Toys, respectively. We attribute these improvements to the high-quality next-item-engagement representations generated by CDDRec. The sequence representation generated from the multi-step denoising process can reveal the user preference from a fine-grained level, thus, being more distinguishable among top-rated candidate items. This approach avoids the *ranking plateau* phenomenon, and obtaining high performance w.r.t top-1 metrics and MRR. On the contrary, baseline methods



(a) Length (b) Frequency (c) Denoising Stage

Fig. 2. The Avg.Change for CDDRec and baseline methods across various subset sequences and denoising stage on Office dataset.

Table 2. Overall Comparison. The best is bolded, and the runner-up is underlined

Dataset	Metric	SVAE	ACVAE	ContrastVAE	CL4Rec	DuoRec	CBiT	GRU4Rec	SASRec	Bert4Rec	FMLP	CDDRec	Imp
Office	R@1	0.0088	0.0139	0.0194	0.0094	0.0120	0.0198	0.0051	0.0198	0.0137	0.0224	**0.0271**	20.98%
	R@5	0.0316	0.0457	0.0642	0.0294	0.0330	0.0593	0.0241	0.0656	0.0485	0.0593	**0.0765**	16.62%
	R@10	0.0597	0.0742	0.1052	0.0430	0.0559	0.0917	0.0510	0.0989	0.0848	0.0901	**0.1091**	3.71%
	N@5	0.0202	0.0300	0.0411	0.0194	0.0223	0.0396	0.0149	0.0428	0.0309	0.0414	**0.0521**	21.73%
	N@10	0.0292	0.0392	0.0544	0.0237	0.0296	0.0500	0.0234	0.0534	0.0426	0.0513	**0.0627**	15.26%
	MRR	0.0249	0.0351	0.0463	0.0207	0.0264	0.0437	0.0204	0.0457	0.0408	0.0455	**0.0548**	18.36%
Beauty	R@1	0.0014	0.0167	0.0161	0.0045	0.0107	0.0174	0.0079	0.0129	0.0119	0.0154	**0.0203**	16.67%
	R@5	0.0068	0.0428	0.0491	0.0160	0.0278	0.0512	0.0266	0.0416	0.0396	0.0433	**0.0542**	5.86%
	R@10	0.0127	0.0606	0.0741	0.0250	0.0403	0.0762	0.0421	0.0633	0.0595	0.0627	**0.0770**	1.05%
	N@5	0.0041	0.0299	0.0327	0.0103	0.0193	0.0343	0.0172	0.0274	0.0257	0.0297	**0.0376**	9.62%
	N@10	0.0060	0.0356	0.0407	0.0131	0.0233	0.0424	0.0222	0.0343	0.0321	0.0360	**0.0447**	5.42%
	MRR	0.0046	0.0310	0.0345	0.0111	0.0201	0.0359	0.0191	0.0291	0.0294	0.0305	**0.0387**	7.80%
Tools	R@1	0.0055	0.0090	0.0108	0.0060	0.0058	0.0066	0.0047	0.0103	0.0059	0.0089	**0.0127**	17.59%
	R@5	0.0118	0.0242	0.0315	0.0189	0.0182	0.0214	0.0154	0.0284	0.0189	0.0251	**0.0359**	13.97%
	R@10	0.0204	0.0364	0.0483	0.0293	0.0361	0.0347	0.0242	0.0427	0.0319	0.0359	**0.0522**	8.07%
	N@5	0.0086	0.0166	0.0212	0.0123	0.0120	0.0139	0.0102	0.0194	0.0123	0.0170	**0.0244**	15.09%
	N@10	0.0114	0.0206	0.0266	0.0156	0.0148	0.0182	0.0129	0.0240	0.0165	0.0204	**0.0297**	11.65%
	MRR	0.0098	0.0178	0.0227	0.0132	0.0128	0.0154	0.0113	0.0207	0.0160	0.0174	**0.0253**	11.45%
Toys	R@1	0.0022	0.0156	0.0228	0.0067	0.0099	0.0195	0.0066	0.0193	0.0110	0.0189	**0.0270**	18.42%
	R@5	0.0057	0.0349	0.0591	0.0180	0.0258	0.0525	0.0226	0.0551	0.0300	0.0516	**0.0665**	12.52%
	R@10	0.0098	0.0492	0.0823	0.0259	0.0360	0.0747	0.0363	0.0797	0.0466	0.0674	**0.0935**	13.61%
	N@5	0.0038	0.0255	0.0414	0.0124	0.0179	0.0364	0.0148	0.0377	0.0206	0.0357	**0.0472**	14.01%
	N@10	0.0038	0.0301	0.0489	0.0149	0.0212	0.0435	0.0192	0.0456	0.0260	0.0408	**0.0559**	14.31%
	MRR	0.0044	0.0270	0.0422	0.0132	0.0182	0.0373	0.0165	0.0385	0.0244	0.0347	**0.0479**	13.51%

encounter *ranking plateau*, where the ratings of top-rate items are indistinguishable. From another perspective, metrics like NDCG@5, NDCG@10, and MRR, which take the ranking position of target items into account, show impressive improvements. This indicates that our model CDDRec ranks target items relatively higher than other models.

4.3 Ablation Study

In this section, we conduct experiments to examine the contributions of the denoising and diffusion phases. Notably, the model is designed to predict the denoised mean of the corrupted target item embedding. By employing this sampling procedure (Eq. 4) with the predicted mean, the model is able to mimic the corrupted target items. Accurate prediction of the mean for these perturbed items, devoid of noise, confers the denoising ability upon the model. On the other hand, the diffusion step also follows a sampling process as described by Eq. 2. To scrutinize the impact of these dual processes, we substitute the two sampling steps with the predicted mean and the original item embedding, respectively, and present the experimental findings in Table 3.

Firstly, the diffusion and denoising processes generally contribute positively to the overall performance across all datasets, as evidenced by the performance decline in comparison to our model CDDRec. Moreover, the significance of diffusion and denoising varies among datasets. Specifically, the denoising process demonstrates greater importance in the Office and Toys datasets, while the diffusion phase is more crucial for the Beauty and Tools datasets, when we compare

the performance drop within each row. An explanation for this variation could be attributed to the differences in sequence length across the datasets. Office and Toys datasets exhibit relatively longer sequences, which could result in a higher likelihood of noisy interactions, thereby rendering the sampling process on the target sequence less effective. Conversely, when dealing with shorter sequences, the diffusion phase that introduces noise may serve as an augmentation strategy, bolstering the model's robustness to noisy interactions.

Table 3. Ablation Study

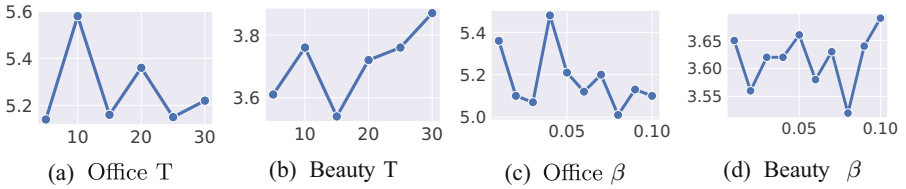
	-Diffusion			-Denoising			CDDRec	
	R@1	MRR	avg.drop	R@1	MRR	avg.drop	R@1	MRR
Office	0.0236	0.0496	11.20%	0.0222	0.0490	14.33%	**0.0271**	**0.0548**
Beauty	0.0154	0.0342	17.88%	0.0179	0.0365	8.75%	**0.0203**	**0.0387**
Tools	0.0112	0.0226	11.24%	0.0116	0.0243	6.31%	**0.0127**	**0.0253**
Toys	0.0267	0.0464	2.12%	0.0264	0.0462	2.89%	**0.0270**	**0.0479**

Table 4. Case Study

Denoising Stage	10	15	20
Rank 1	Swivel Tower Sorter	Swivel Tower Sorter	**Wristbands**
Rank 2	Desk Tray	Paper Clip Holder	Erase Markers
Rank 3	Paper Clip Holder	Desk Sorter	Pencil Sharpener
Rank 4	Desk Sorter	Desk Tray	Graphite Pencils
Rank of target item	>40	7	1

4.4 Hyperparameter Sensitivity

In this section, we investigate the sensitivities of CDDRec's hyperparameters. Due to space constraints, we focus on reporting the experimental results for key hyperparameters, including the maximum diffusion step and maximum noise schedule. As depicted in Fig. 3, the optimal maximum diffusion steps are 10 and 30 for Office and Beauty datasets, respectively. The possible reason is that the Beauty dataset has a higher number of items, presenting a greater challenge for CDDRec to learn meaningful item embeddings. Consequently, a greater number of denoising steps is required to refine intermediate states of item representations. We observe that the optimal maximum noise levels for Office and Beauty are 0.04 and 0.1, respectively. A possible explanation is that longer sequences may inherently contain noisy interactions, thus necessitating less added noise.



(a) Office T (b) Beauty T (c) Office β (d) Beauty β

Fig. 3. The evaluation of CDDRec on MRR through two datasets with different maximum diffusion step T and noise schedule β.

4.5 Case Study for Stepwise Generation

In the effort to demystify the intermediate generation of CDDRec, we execute item retrieval utilizing the intermediate-generated preference and item representations. The top-4 retrieved items along with the rank of the actual target

item are presented in Table 4. Prior interacted items encompass *Desk Organizer, Organization Cube, Pencil Cup, Wristbands*, with the target item also being wristbands. A notable observation from history is that the sequence is dominated by collection-related kits, with a sudden shift in user engagement towards wristbands. After the initial ten denoising stages, the system continues to suggest collection kits, and the target item's rank falls outside the top 40. As the system undergoes more stepwise denoising, CDDRec begins to acknowledge the significance of the most recent purchase behavior, subsequently improving the ranking of the target items. After 20 stages of denoising, CDDRec manages to place the target item at the top position, and the recommended items display greater diversity, including items such as *pencil sharpener, graphite pencil*, etc.

5 Conclusion

In summary, we highlight the *ranking plateau* issue and underline the importance of stepwise generation as an effective solution. We introduce CDDRec, a model characterized by a cross-attentive conditional denoising decoder. This decoder makes use of the denoising step indicator and sequence encoder output as input, predicting the denoised mean at each denoising step. We also propose the cross-divergence objective with contrastive loss, tailored for sequence recommendation. These objectives guard against representation collapse while enabling the model to exhibit ranking capacity. Consequently, CDDRec can generate high-fidelity sequence/item representations and provide fine-grained ranking predictions, thus addressing *ranking plateau* issues. Thorough experimental results indicate CDDRec's superior performance, outshining contemporary SOTA methods, especially in top-1 metrics.

Acknowledgement. This work is supported in part by NSF under grant III-2106758.

References

1. Du, H., et al.: Contrastive learning with bidirectional transformers for sequential recommendation. In: CIKM (2022)
2. Du, H., Yuan, H., Huang, Z., Zhao, P., Zhou, X.: Sequential recommendation with diffusion models. arXiv preprint arXiv:2304.04541 (2023)
3. Gao, Z., et al.: Difformer: empowering diffusion model on embedding space for text generation. arXiv preprint arXiv:2212.09412 (2022)
4. Gong, S., Li, M., Feng, J., Wu, Z., Kong, L.: DiffuSeq: sequence to sequence text generation with diffusion models. arXiv preprint arXiv:2210.08933 (2022)
5. Guo, X., Wang, Y., Du, T., Wang, Y.: ContraNorm: a contrastive learning perspective on oversmoothing and beyond. arXiv preprint arXiv:2303.06562 (2023)
6. He, J., Cheng, L., Fang, C., Zhang, D., Wang, Z., Chen, W.: Mitigating undisciplined over-smoothing in transformer for weakly supervised semantic segmentation. arXiv preprint arXiv:2305.03112 (2023)

7. Hidasi, B., Karatzoglou, A.: Recurrent neural networks with top-k gains for session-based recommendations. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, 22–26 October 2018, pp. 843–852. ACM (2018)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
9. Huang, C.W., Lim, J.H., Courville, A.C.: A variational perspective on diffusion-based generative models and score matching. Adv. Neural. Inf. Process. Syst. **34**, 22863–22876 (2021)
10. Kang, W., McAuley, J.J.: Self-attentive sequential recommendation. In: IEEE International Conference on Data Mining, ICDM 2018, Singapore, 17–20 November 2018, pp. 197–206. IEEE Computer Society (2018)
11. Li, X., Thickstun, J., Gulrajani, I., Liang, P.S., Hashimoto, T.B.: Diffusion-LM improves controllable text generation. Adv. Neural. Inf. Process. Syst. **35**, 4328–4343 (2022)
12. Li, Z., Sun, A., Li, C.: DiffuRec: a diffusion model for sequential recommendation. arXiv preprint arXiv:2304.00686 (2023)
13. Liu, Z., Fan, Z., Wang, Y., Yu, P.S.: Augmenting sequential recommendation with pseudo-prior items via reversely pre-training transformer. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1608–1612 (2021)
14. McAuley, J.J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015, pp. 43–52. ACM (2015)
15. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR (2021)
16. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
17. Qiu, R., Huang, Z., Yin, H., Wang, Z.: Contrastive learning for representation degeneration problem in sequential recommendation. In: WSDM 2022: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event/Tempe, AZ, USA, 21–25 February 2022, pp. 813–823. ACM (2022)
18. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
19. Rasul, K., Seward, C., Schuster, I., Vollgraf, R.: Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In: International Conference on Machine Learning, pp. 8857–8868. PMLR (2021)
20. Sachdeva, N., Manco, G., Ritacco, E., Pudi, V.: Sequential variational autoencoders for collaborative filtering. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, 11–15 February 2019, pp. 600–608. ACM (2019)
21. Savinov, N., Chung, J., Binkowski, M., Elsen, E., Oord, A.v.d.: Step-unrolled denoising autoencoders for text generation. arXiv preprint arXiv:2112.06749 (2021)
22. Sun, F., et al.: BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, 3–7 November 2019, pp. 1441–1450. ACM (2019)
23. Wang, W., Xu, Y., Feng, F., Lin, X., He, X., Chua, T.S.: Diffusion recommender model. arXiv preprint arXiv:2304.04971 (2023)

24. Wang, Y., Liu, Z., Zhang, J., Yao, W., Heinecke, S., Yu, P.S.: DRDT: dynamic reflection with divergent thinking for LLM-based sequential recommendation. arXiv preprint arXiv:2312.11336 (2023)
25. Wang, Y., et al.: Exploiting intent evolution in e-commercial query recommendation. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 5162–5173 (2023)
26. Wang, Y., Zhang, H., Liu, Z., Yang, L., Yu, P.S.: ContrastVAE: contrastive variational autoencoder for sequential recommendation. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 2056–2066 (2022)
27. Xie, X., et al.: Contrastive learning for sequential recommendation. arXiv preprint arXiv:2010.14395 (2020)
28. Xie, Z., Liu, C., Zhang, Y., Lu, H., Wang, D., Ding, Y.: Adversarial and contrastive variational autoencoder for sequential recommendation. In: WWW 2021: The Web Conference 2021, Virtual Event/Ljubljana, Slovenia, 19–23 April 2021, pp. 449–459. ACM/ IW3C2 (2021)
29. Zhou, K., et al.: S3-Rec: self-supervised learning for sequential recommendation with mutual information maximization. In: CIKM 2020: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, 19–23 October 2020, pp. 1893–1902. ACM (2020)
30. Zhou, K., Yu, H., Zhao, W.X., Wen, J.R.: Filter-enhanced MLP is all you need for sequential recommendation. In: Proceedings of the ACM Web Conference 2022, pp. 2388–2399 (2022)