Addressing Cold-Start Problem in Click-Through Rate Prediction via Supervised Diffusion Modeling

Wenqiao Zhu, Lulu Wang, Jun Wu

HiThink Research
{zhuwenqiao, wanglulu2, wujun2}@myhexin.com

Abstract

Predicting Click-Through Rates is a crucial function within recommendation and advertising platforms, as the output of CTR prediction determines the order of items shown to users. The Embedding and MLP paradigm has become a standard approach for industrial recommendation systems and has been widely deployed. However, this paradigm suffers from cold-start problems, where there is either no or only limited user action data available, leading to poorly learned ID embeddings. The cold-start problem hampers the performance of new items. To address this problem, we design a novel diffusion model to generate a warmed-up embedding for new items. Specifically, we define a novel diffusion process between the ID embedding space and the side information space. In addition, we can derive a sub-sequence from the diffusion steps to expedite training, given that our diffusion model is non-Markovian. Our diffusion model is supervised by both the variational inference and binary cross-entropy objectives, enabling it to generate warmed-up embeddings for items in both the cold-start and warm-up phases. Additionally, we have conducted extensive experiments on three recommendation datasets. The results confirmed the effectiveness of our approach.

Code — https://github.com/WNQzhu/CSDM

Introduction

Recommendation systems are crucial components of numerous commercial platforms, addressing the challenge of information overload prevalent in the digital age. A primary goal of many such systems often involves predicting Click-Through Rates (CTR). Embedding & MLP (Multilayer Perceptron) methods (Guo et al. 2017, 2021; Zhou et al. 2019) have been widely utilized for this purpose. However, the Embedding & MLP approach faces the cold-start problem, which arises from the long-tail distribution of candidate items and the dynamic nature of real-world recommendation systems. New items in recommendation systems often have no or very limited user interactions. Consequently, the embeddings for these items are not adequately learned, leading to sub-optimal performance in predicting their CTRs. Furthermore, the embedding layer makes up a substantial part of the model's parameters and determines the input for the feature interaction and MLP modules. Thus, it is imperative to improve the embedding layer to mitigate the cold-start problem in CTR prediction tasks.

To address the challenge of the cold-start problem, several approaches have been proposed that leverage either the limited available samples or the side information associated with new items. Approaches that make use of the limited available samples often employ a meta-learning framework, such as MAML (Finn, Abbeel, and Levine 2017), to derive robust representations for new items through an optimized training procedure (Pan et al. 2019; Lee et al. 2019; Lu, Fang, and Shi 2020; Ouyang et al. 2021; Zhu et al. 2021). On the other hand, methods that leverage side information generally learn a transformation from user/item attributes to robust embeddings (Mo et al. 2015; Roy and Guntuku 2016; Saveski and Mantrach 2014; Schein et al. 2002; Seroussi, Bohnert, and Zukerman 2011; Wei et al. 2021; Zhu et al. 2020).

The aforementioned approaches learn representations as fixed points in the embedding space. However, due to the limited data available in cold-start scenarios, it is highly challenging to learn a reliable representation (Zhang et al. 2019). To achieve reliable embedding learning for cold-start items, some variational approaches have been proposed (Xu et al. 2022; Zhao et al. 2022). These approaches treat embedding learning as a distribution estimation problem and have shown effectiveness in tackling the cold-start problem. However, these approaches struggle with the trade-off between traceability and flexibility (Kingma et al. 2016; Sohl-Dickstein et al. 2015; Wang et al. 2023), and they also suffer from the model collapse problem (Shi et al. 2020). To address the limitations of existing methods and further improve the performance of **c**old-**s**tart in CTR prediction, we propose a **d**iffusion **m**odel named CSDM, which constructs the transition between embeddings and side information in a denoising manner. While diffusion models (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015) have achieved remarkable results in image synthesis tasks (Teng et al. 2024) and can address the limitations in variational auto-encoder methods, it is not straightforward to adopt diffusion models to tackle the cold-start problem directly due to the following reasons: (1) Standard diffusion models construct transitions between the data distribution and standard Gaussian distri-

bution; in the context of the cold-start problem, we need to construct transitions between embeddings and side information. (2) Diffusion models require more training and inference time due to the lengthy denoising steps.

We propose a novel diffusion model to overcome the aforementioned obstacles in addressing the cold-start problem. Specifically, in addition to gradually adding noise during the forward process of diffusion steps, we also progressively incorporate portions of side information. This approach enables us to construct a transition between ID embeddings and side information. We consider our model as a non-Markovian model, which permits us to extract a sub-sequence during the generation process to expedite the training speed. We update the original embeddings with the newly generated ones, ensuring no additional inference cost is incurred during the inference phase. We perform extensive experiments on three CTR prediction benchmark datasets to validate the effectiveness of our proposed method.

In a nutshell, the contributions of this work include:

- We propose a diffusion model to address the cold-start problem in CTR prediction tasks while considering both training and inference costs. To the best of our knowledge, we are the first to employ diffusion models for cold-start problems in CTR predictions.
- We design a new diffusion process that allows us to construct transitions between ID embeddings and side information. Furthermore, our model is non-Markovian, which enables us to extract sub-sequences to reduce training costs.
- Extensive experiments are conducted on three public benchmark datasets, and the results show that CSDM outperforms existing cold-start methods in CTR predictions.

Preliminary: Diffusion Models

Diffusion models represent a class of generative models that leverage the diffusion process to remove noise from latent samples incrementally, resulting in the generation of new samples. DDPM (Ho, Jain, and Abbeel 2020) is one of the most representative diffusion models, comprising both a forward process and a reverse process.

Forward Process gradually adds noises to the given original data $\mathbf{z}_0 \in R^d$ over a sequence of T steps, creating a Markov chain $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_T$. In this chain, $\mathbf{z}_T$ is assumed to be an approximation of Gaussian noise. Each forward step is defined as:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) := \mathcal{N}\left(\mathbf{z}_t; \sqrt{\frac{\alpha_t}{\alpha_{t-1}}}\mathbf{z}_{t-1}, \left(1 - \frac{\alpha_t}{\alpha_{t-1}}\right)\mathbf{I}\right).$$

DDPM incorporates a pre-established, constant noise schedule $\alpha_{1:T} \in (0, 1]^T$, which controls the quantity of noise introduced at each step. It admits a closed form of $\mathbf{z}_t$ at any timestep t: $q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\alpha_t}\mathbf{z}_0, (1 - \alpha_t)\mathbf{I})$.

Reverse Process sequentially eliminates the noise from $\mathbf{z}_t$ to recover $\mathbf{z}_{t-1}$ in a recursive manner, continuing this process until it reaches the initial step 0. A single transition step parameterized by $p_\omega(\mathbf{z}_{t-1}|\mathbf{z}_t) := \mathcal{N}(\mathbf{z}_{t-1}; \mu_\omega(\mathbf{z}_t, t), \mathbf{\Sigma}_\omega(\mathbf{z}_t, t))$ is learned, where $\mu_\omega(\mathbf{z}_t, t)$ and $\mathbf{\Sigma}_\omega(\mathbf{z}_t, t)$ are learned mean and variance. A U-net (Ronneberger, Fischer, and Brox 2015) architecture is employed to model these parameters.

Optimization. Given the definition of the forward and reverse process, DDPM optimizes the Evidence Lower Bound (ELBO) objective function. It calculates the KL-divergence between p_ω and q plus an entropy term:

$$\begin{aligned}\mathcal{L}_{\text{diff}} = \mathbb{E}_q &\left[\underbrace{D_{KL}(q(\mathbf{z}_t|\mathbf{z}_0)\|p(\mathbf{z}_T))}_{\mathcal{L}_T}\right] \\ &+ \mathbb{E}_q\left[\sum_{t>1}\underbrace{D_{KL}(q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0)\|p_\omega(\mathbf{z}_{t-1}|\mathbf{z}_t))}_{\mathcal{L}_{t-1}}\right] \\ &- \mathbb{E}_q\left[\underbrace{\log p_\omega(\mathbf{z}_0|\mathbf{z}_1)}_{\mathcal{L}_0}\right] \end{aligned} \tag{1}$$

An inherent limitation of DDPM is its dependence on a Markovian process, which leads to a computationally intensive reverse process. To address this limitation, DDIM (Song, Meng, and Ermon 2020) adopts a non-Markovian method, markedly accelerating the reverse process. The objective function of DDIM is equivalent to Equation (1), up to a constant difference.

Method

Problem Definition

The objective of Click-Through Rate (CTR) prediction is to forecast the likelihood that a user will click on a specific presented item. The outcome of this prediction will determine the final ordering of items shown to users. A CTR prediction task is typically structured as a supervised binary classification problem and is trained using an i.i.d (independent and identically distributed) dataset $\mathcal{D}$ from users' historical interactions. Each instance $(\mathbf{x}, y) \in \mathcal{D}$ includes a collection of features $\mathbf{x}$ and a target label $y \in \{0, 1\}$, indicating the user's reaction to the presented item. Generally, the input feature $\mathbf{x}$ can be decomposed into several components:

$$\mathbf{x} = [u, \mathcal{X}_u, i, \mathcal{X}_i, \textit{context}] \tag{2}$$

Here, u is a unique identifier for each user within the recommendation system. Similarly, i is a unique identifier for each item. $\mathcal{X}_u$ represents the set of features associated with users, while $\mathcal{X}_i$ represents the set of features with items. The *context* refers to environmental features such as time and location.

The Embedding & MLP paradigm first employs embedding technology to convert the IDs and features into unique embeddings. We designate the embeddings for the item ID, use ID, item features, user features, and context features as $\mathbf{e}_i \in R^d$, $\mathbf{e}_u \in R^d$, $\mathbf{e}_{\mathcal{X}_i} \in R^{d\times|\mathcal{X}_i|}$, $\mathbf{e}_{\mathcal{X}_u} \in R^{d\times|\mathcal{X}_u|}$, $\mathbf{e}_c \in R^{d\times|c|}$, respectively. Here, d is the dimension of the embeddings. The CTR model estimates the probability $\hat{y} = Pr(y = 1|\mathbf{x})$ by applying a discriminative function $f(\cdot)$:

$$\hat{y} = f(\mathbf{e}_i, \mathbf{e}_u, \mathbf{e}_{\mathcal{X}_i}, \mathbf{e}_{\mathcal{X}_u}, \mathbf{e}_c; \theta, \phi) \tag{3}$$

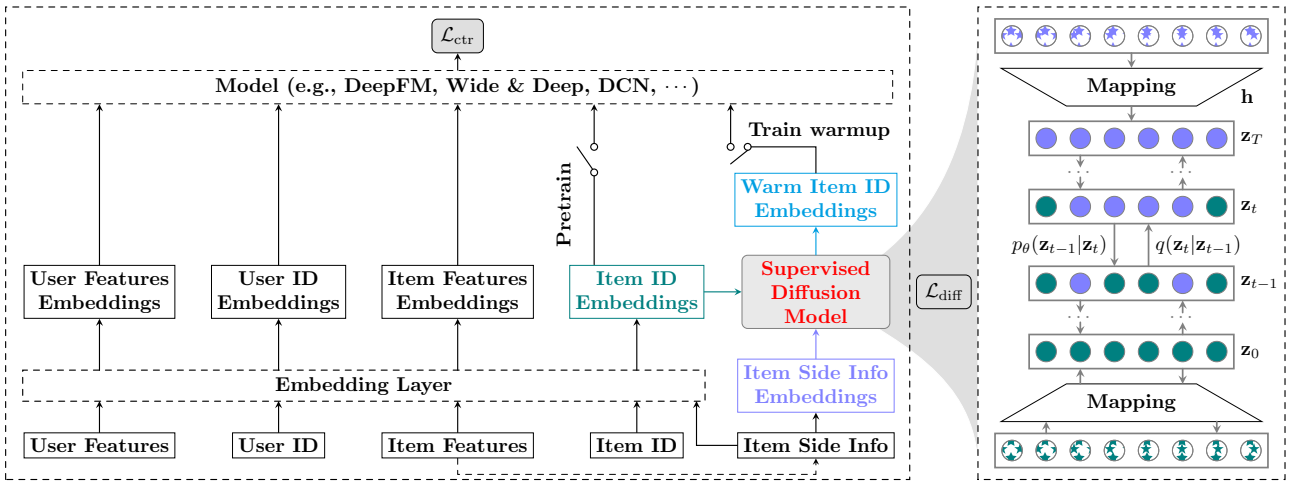

Figure 1: The proposed CSDM framework for cold-start problems in CTR prediction.

where θ denotes the parameters of the backbone deep model $f(\cdot)$ and ϕ denotes the parameters of the embedding layers. The Binary Cross Entropy is often employed as the loss function for binary classification:

$$\mathcal{L}_{ctr}(\theta, \phi) = -y \log \hat{y} - (1-y) \log(1-\hat{y}) \tag{4}$$

Since the parameters are trained by users' historical behavior data, the embedding $\mathbf{e}_i$ for recently emerged items $(i, \mathcal{X}_i)$ with limited user interactions are not well learned. As a result, the backbone model $f(\cdot)$ struggles to make an accurate estimation of the probability for these items. This issue is known as the item cold-start problem, which includes two stages: (1) The cold-start phase, during which there are no user interactions with the item, and (2) the warm-up phase, where there are a limited number of user interactions. In this work, we tackle both of these stages, focusing exclusively on the item cold-start problem. Specifically, a subset of item features $\mathcal{X}_v \subset \mathcal{X}_i$ are employed with our diffusion model to address the cold-start problem in CTR prediction.

Supervised Diffusion Model

To address the cold-start problem, we utilize a subset of item features $\mathcal{X}_v$, referred to as side information, to enrich the item ID embeddings. To achieve this, we employ a diffusion model to enable the flow of semantic information between the side information embeddings and the ID embeddings. However, since the standard diffusion process transforms an embedding into Gaussian noise, it is not directly applicable in this context. Therefore, we design a new diffusion process to learn warm-up ID embeddings for items, as illustrated in Figure 1.

We first pre-train a backbone model to provide the initial item ID embeddings. Then, we conduct the diffusion process to generate warmed-up embeddings. Following (Li et al. 2022; Cui et al. 2024), we convert the discrete side information to a continuous space using an embedding map and project the initial item ID embeddings into hidden states. Let $\mathbf{h}$ and $\mathbf{z}_0$ denote the hidden state of the side information and the initial embedding, respectively. We gradually transform $\mathbf{z}_0$ into $\mathbf{h}$ using a forward diffusion process:

$$\mathbf{z}_0 \to \mathbf{z}_1 \to \cdots \to \mathbf{z}_T = \mathbf{h} + \epsilon, \tag{5}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since our diffusion process depends on both $\mathbf{z}_0$ and $\mathbf{h}$, it is no longer Markovian. Inspired by DDIM (Song, Meng, and Ermon 2020), we define a family $\mathcal{Q}$ of inference distributions, indexed by a real vector $\sigma \in \mathbb{R}^T_{\geq 0}$:

$$q_\sigma(\mathbf{z}_{1:T}|\mathbf{z}_0, \mathbf{h}) := q_\sigma(\mathbf{z}_T|\mathbf{z}_0, \mathbf{h}) \prod_{t=2}^{T} q_\sigma\left(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{h}, \mathbf{z}_0\right)$$

where

$$q_\sigma(\mathbf{z}_T|\mathbf{z}_0, \mathbf{h}) = \mathcal{N}\left(\sqrt{\alpha_T}\mathbf{z}_0 + \sqrt{c_T}\mathbf{h}, (1-\alpha_T)\mathbf{I}\right) \tag{6}$$

The forward process is governed by a decreasing sequence $\alpha_t \in (0,1]^T$ and an increasing sequence $c_t \in (0,1]^T$. In ideal case, when $\alpha_t \to 0$ and $c_t \to 1$, we have $\mathbf{z}_T \sim \mathcal{N}(\mathbf{h}, \mathbf{I})$. In the recommendation scenario, since $\mathbf{z}_t$ contains collaborative filtering information and $\mathbf{h}$ contains feature information, it is not necessary to fully zero out $\mathbf{z}_0$ or $\mathbf{h}$. Furthermore, we choose the mean function as

$$q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}_0, \mathbf{h}) = \mathcal{N}(\mathbf{z}_{t-1}|\kappa_t\mathbf{z}_t + \lambda_t\mathbf{z}_0 + \nu_t\mathbf{h}, \sigma_t^2\mathbf{I})$$

in order to guarantee that the following equation

$$q_\sigma(\mathbf{z}_t|\mathbf{z}_0, \mathbf{h}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{c_t}\mathbf{h}, (1-\alpha_t)\mathbf{I}) \tag{7}$$

holds true for all t. The parameters are set as:

$$\kappa_t = \sqrt{\frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}}$$

$$\lambda_t = \sqrt{\alpha_{t-1}} - \sqrt{\alpha_t}\sqrt{\frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}}$$

$$\nu_t = \sqrt{c_{t-1}} - \sqrt{c_t}\sqrt{\frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}}$$

This posterior function can be derived from Bayes' rule, we show the proof in the supplementary materials.

In the reverse process, given $\mathbf{z}_t$, we can predict the denoised observation of the hidden state of ID embeddings:

$$g_\omega^{(t)}(\mathbf{z}_t) := (\mathbf{z}_t - \sqrt{c_t}\mathbf{h} - \sqrt{1-\alpha_t}\epsilon_\omega^t(\mathbf{z}_t))/\sqrt{\alpha_t} \quad (8)$$

where $\{\epsilon_\omega^{(t)}\}_{t=1}^T$ is a set of T functions to predict noise from $\mathbf{z}_t$ and ω contains the learnable parameters of diffusion model. Therefore, we can define the reserve step as:

$$p_\omega^{(t)}(\mathbf{z}_{t-1}|\mathbf{z}_t) = \begin{cases} g_\omega^{(1)}(\mathbf{z}_t) & \text{if } t=1; \\ q_\sigma(\mathbf{z}_{t-1}|\mathbf{z}_t, g_\omega^{(t)}(\mathbf{z}_t), \mathbf{h}) & \text{otherwise} \end{cases}$$

Optimization. The parameters of diffusion models are optimized by a combination of the $\mathcal{L}_{\text{ctr}}$ and the variational inference objective $\mathcal{L}_{\text{diff}}$:

$$\mathcal{L} = \mathcal{L}_{\text{ctr}} + \rho\mathcal{L}_{\text{diff}}, \quad (9)$$

where ρ is a hyper-parameter. By combining $\mathcal{L}_{\text{ctr}}$ and $\mathcal{L}_{\text{diff}}$, the generative process of the diffusion model takes into account both the collaborative filtering information derived from user action data and the side information obtained from item features. Consequently, our method is applicable in both the cold-start and warm-up stages of the CTR prediction task. In our implementation, we utilize the simplified version of $\mathcal{L}_{\text{diff}}$, as proposed in DDPM (Ho, Jain, and Abbeel 2020).

Generating Warmed-up Embeddings. Given the reverse step provided above, we can generate the warmed-up embeddings from $\mathbf{h}$ by repeating the following step:

$$\begin{aligned} \mathbf{z}_{t-1} = \sqrt{\alpha_{t-1}} & \underbrace{\left(\frac{\mathbf{z}_t - \sqrt{c_t}\mathbf{h} - \sqrt{1-\alpha_t}\epsilon_\omega^{(t)}(\mathbf{z}_t)}{\sqrt{\alpha_t}}\right)}_{\text{predicted hidden state of initial ID embedding}} \\ & + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\omega^{(t)}(\mathbf{z}_t)}_{\text{direction pointing to } \mathbf{z}_t} \\ & + \sqrt{c_{t-1}} \underbrace{\mathbf{h}}_{\text{side information}} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}} \end{aligned} \quad (10)$$

From Equation (10), we can observe that the generating process consists of two key components: (1) predicting the relevant ID embedding, and (2) leveraging side information to improve the generation quality. In this process, we set $\sigma_t = 0$ for all t to ensure that the reverse process is deterministic. Upon obtaining the estimated warmed-up embedding $\tilde{\mathbf{z}}_0$, we project it back into the ID embedding space via a linear function. Following the approach taken by CVAR (Zhao et al. 2022), we incorporate the frequency of items into this linear function. The original ID embeddings are subsequently replaced with the newly generated ones. As a result, our method does not modify the backbone structures and incurs no extra computational overhead during the inference phase.

Sub-sequence. Our approach is non-Markovian. Therefore, we can consider a sub-sequence of the latent variables $\mathbf{z}_{1:T}$ to accelerate the generative process. We uniformly sample a sub-sequence from $\mathbf{z}_{1:T}$ using a step parameter s. A larger s results in a smaller sub-sequence, which in turn enables a faster generative process.

Experiment

Dataset

We evaluate our method on three publicly available datasets: **MovieLens-1M** [1], **Taobao Display Ad Click** [2], and **CIKM 2019 EComm AI** [3]. The details of these datasets are described in the supplementary material.

Baselines

We compare our method with two groups of Click-Through Rate (CTR) prediction methods. The first group encompasses a variety of common feature-crossing techniques tailored specifically for CTR prediction. Methods within this group also serve as the foundational model for numerous cold-start algorithms. (1) DeepFM (Guo et al. 2017) is a method that combines low-order feature interactions through Factorization Machines (FM) (Rendle 2010) and high-order feature interactions through a deep neural network. (2) Wide & Deep (Cheng et al. 2016) contains a linear model and a deep neural network to effectively handle both simple and complex relationships in the data. (3) DCN (Wang et al. 2017) explicitly models the interactions between features using a deep network.

The other group comprises state-of-the-art methods aimed at addressing the cold-start problem in CTR prediction tasks. We compare our method with the following methods: DropoutNet (Volkovs, Yu, and Poutanen 2017), MWUF (Zhu et al. 2021), Meta-E (Pan et al. 2019), VELF (Xu et al. 2022), and CVAR (Zhao et al. 2022).

Experimental Settings

Dataset Splits. We divided the datasets into several groups following (Zhu et al. 2021) to assess the performance of our proposed method in both the cold-start and warm-up phases. First, we divide the items into two groups based on their frequency using a threshold N. Items with a frequency greater than N are classified as old items, while those with a lower frequency are considered new items. The threshold N is set to 200 for MovieLens-1M, 2000 for TaobaoAD, and 200 for CIKM 2019, ensuring the ratio of new items to old items is approximately 8:2, which mirrors a long-tail distribution as described in (Chen et al. 2020). We further divide the new item instances, sorted by timestamps, into four groups: warm-a, warm-b, warm-c, and a test set. The first $3 \times K$ instances are distributed evenly among warm-a, warm-b, and warm-c, with the remainder allocated to the test set. The value of K is set to 20 for MovieLens-1M, 500 for TaobaoAD, and 50 for CIKM 2019, respectively.

Implementation Details. We apply consistent experimental settings across all methods for each dataset to ensure fair comparisons. The embedding size for all features is set to 16 for compared methods. Additionally, the MLPs in the backbone models utilize two dense layers, each with 16 units. We set the learning rate to 0.001 for all methods, and the mini-batch size is set to 2048 for MovieLens-1M and TaobaoAD,

[1] http://www.grouplens.org/datasets/movielens/

[2] https://tianchi.aliyun.com/dataset/dataDetail?dataId=56

[3] https://tianchi.aliyun.com/competition/entrance/231721

	Methods	Cold		Warm-a		Warm-b		Warm-c	
		AUC	RelaImpr	AUC	RelaImpr	AUC	RelaImpr	AUC	RelaImpr
MovieLens-1M	DeepFM	0.7313	0.00%	0.7464	0.00%	0.7588	0.00%	0.7692	0.00%
	DropoutNet(DeepFM)	0.7410	4.19%	0.7506	1.70%	0.7593	0.19%	0.7671	-0.78%
	MWUF(DeepFM)	0.7324	0.47%	0.7466	0.08%	0.7590	0.07%	0.7694	0.07%
	Meta-E(DeepFM)	0.7397	3.63%	0.7513	1.98%	0.7614	1.00%	0.7690	-0.07%
	VELF(DeepFM)	0.7244	-2.98%	0.7695	9.37%	0.7382	-7.95%	0.7741	1.82%
	CVAR(DeepFM)	0.7349	1.55%	0.7910	18.10%	0.8009	16.27%	0.8044	13.07%
	CSDM(DeepFM)	**0.7443**	**5.62%**	**0.7982**	**21.02%**	**0.8058**	**18.16%**	**0.8089**	**14.74%**
	Methods	Cold		Warm-a		Warm-b		Warm-c	
		AUC	RelaImpr	AUC	RelaImpr	AUC	RelaImpr	AUC	RelaImpr
Taobao AD	DeepFM	0.5958	0.00%	0.6089	0.00%	0.6204	0.00%	0.6306	0.00%
	DropoutNet(DeepFM)	0.5970	1.25%	0.6097	0.73%	0.6207	0.25%	0.6305	-0.7%
	MWUF(DeepFM)	0.5967	0.94%	0.6101	1.10%	0.6207	0.25%	0.6303	-0.23%
	Meta-E(DeepFM)	0.5975	1.77%	0.6119	2.75%	0.6226	1.83%	0.6323	1.30%
	VELF(DeepFM)	0.5967	0.93%	0.6176	7.98%	0.6258	4.49%	0.6335	2.22%
	CVAR(DeepFM)	0.5998	4.17%	0.6194	9.64%	0.6295	7.56%	0.6370	4.90%
	CSDM(DeepFM)	**0.6004**	**4.80%**	**0.6290**	**18.45%**	**0.6324**	**9.97%**	**0.6382**	**5.82%**
	Methods	Cold		Warm-a		Warm-b		Warm-c	
		AUC	RelaImpr	AUC	RelaImpr	AUC	RelaImpr	AUC	RelaImpr
CIKM 2019	DeepFM	0.7376	0.00%	0.7522	0.00%	0.7605	0.00%	0.7671	0.00%
	DropoutNet(DeepFM)	0.7367	-0.38%	0.7487	-1.39%	0.7569	-1.38%	0.7636	-1.31%
	MWUF(DeepFM)	0.7372	-0.17%	0.7501	-0.83%	0.7598	-0.27%	0.7674	0.11%
	Meta-E(DeepFM)	0.7367	-0.38%	0.7483	-1.55%	0.7574	-1.19%	0.7651	-0.75%
	VELF(DeepFM)	0.7403	1.13%	0.7393	-5.11%	0.7390	-8.25%	0.7317	-13.25%
	CVAR(DeepFM)	0.7405	1.22%	0.7588	2.62%	0.7649	1.69%	0.7687	0.60%
	CSDM(DeepFM)	**0.7418**	**1.77%**	**0.7624**	**4.04%**	**0.7686**	**3.10%**	**0.7710**	**1.46%**

Table 1: Model comparison on three datasets. DeepFM is utilized as the backbone. Ten runs are conducted for each method. The best and second-best improvements are highlighted in bold and underlined, respectively.

Split \ ρ	10^{-3}	10^{-2}	10^{-1}	1
Cold	0.5861	0.5875	0.5933	0.6001
Warm-a	0.6305	0.6307	0.6305	0.6296
Warm-b	0.6329	0.6331	0.6329	0.6325
Warm-c	0.6379	0.6384	0.6383	0.6380

Table 2: Performance evaluation using AUC metric on the TaobaoAD dataset split using DeepFM as the backbone model across a range of ρ values, mean of three runs.

and 4096 for CIKM 2019. All methods are optimized using the Adam optimizer (Kingma and Ba 2015) on shuffled samples. We set $T = 100$ for the total number of forward steps. The parameters ρ and s are searched from the sets $\{0.001, 0.01, 0.1, 1\}$ and $\{5, 10\}$, respectively. We define the sequences $\{\alpha_t\}$ and $\{c_t\}$ using a hyper-parameter $\beta = 10^{-5}$ for all experiments:

$$\alpha_t = (1 - \beta)^t \tag{11}$$

$$c_t = \left(\sum_{k=1}^{t} \sqrt{\frac{\alpha_t}{\alpha_k}}\right) / \left(\sum_{k=1}^{T} \sqrt{\frac{\alpha_t}{\alpha_k}}\right) \tag{12}$$

Clearly, $\{\alpha_t\}$ is a decreasing sequence and $\{c_t\}$ is an increasing sequence. We employ position encoding, as described in (Vaswani et al. 2017), integrated into the U-Net to identify the generation steps. Furthermore, to avoid overfitting in the diffusion model, we adopt dropout with $p = 0.5$ in the U-Net. Specifically, we have

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{c_t}\text{drop}(\mathbf{h}, p = 0.5) + \sqrt{1 - \alpha_t}\epsilon \tag{13}$$

in the forward process, where $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$. It's worth noting that the dropout utilized here differs from DropoutNet in that Dropout affects the backbone model, whereas ours impacts the diffusion model only.

Evaluation Metrics. We use the Area Under the Curve (AUC) (Ling, Huang, and Zhang 2003) as the metric to evaluate performance. This is a widely used metric in both recommendation systems (Tang et al. 2020) and computational advertising (Zhou et al. 2018; Pan et al. 2019). An AUC value of 0.5 corresponds to random guessing. (Yan and Li 2014) proposed a relative improvement (RelaImpr) metric to assess the performance improvement, which is calculated

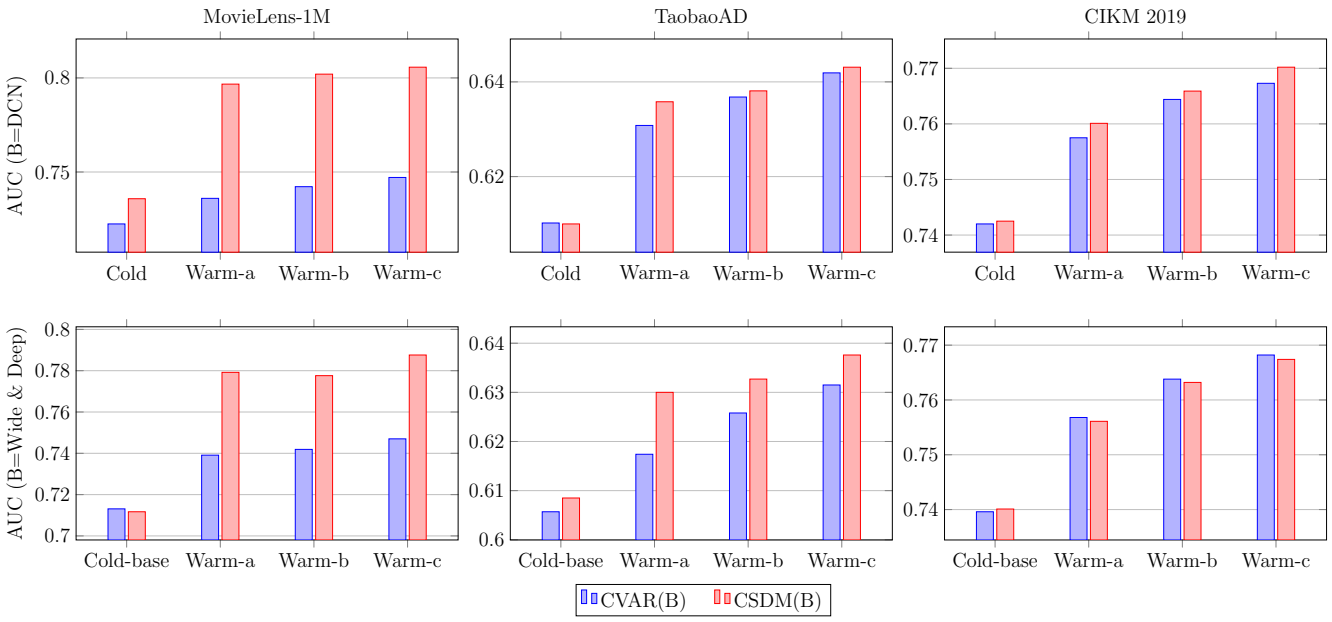

Figure 2: AUC scores evaluated across various stages for different backbone models, conducted over three datasets with 10 runs per model.

as follows:

$$\text{RelaImpr} = \left(\frac{\text{AUC(model)} - 0.5}{\text{AUC(baseline)} - 0.5} - 1 \right) \times 100\% \quad (14)$$

We use this metric to compare the relative improvement in performance across all methods.

Experiment Results

Comparison with State-of-the-Arts. Our approach refines the embeddings of cold item IDs through a diffusion process, all without altering the underlying model architecture. Consequently, we compare our CSDM method against a range of state-of-the-art cold-start solutions from an embedding learning perspective, encompassing DropoutNet, MWUF, Meta-E, VELF, and CVAR. Simultaneously, we select DeepFM, a well-known method for CTR prediction, as the base model. We assess the average outcomes across ten runs on three distinct datasets and present these results in Table 1.

We can notice that CSDM outperforms other comparative baselines across all datasets. The results indicate that our supervised diffusion model is capable of learning high-quality initial embeddings. Apart from learning better initial embeddings, our approach also demonstrates enhanced utilization of user action data during the warm-up phase. This is corroborated by the noticeable performance gains observed in the warm-up stage, as shown in Table 1.

In addition to the aforementioned observations, we have noticed a decline in the relative improvement as items reach maturity. This trend can be attributed to the growing influence of user action data as items garner more interactions. The progressive increase in AUC scores from Warm-a to Warm-b, and Warm-c stages underscores this trend. In later stages, the backbone model is also capable of learning more refined embeddings, given that these items have been interacted with by a larger user base.

Generalization Experiments. Our approach refines the ID embeddings to tackle the cold-start challenge in CTR prediction, making it model-agnostic. We demonstrate its versatility by performing experiments on a range of backbone models beyond DeepFM, including Wide & Deep and DCN. More experiments are provided in the supplementary material. For each model variant, we execute 10 trials to obtain the average AUC across three datasets, with the results reported in Figure 2.

Upon analyzing the results in Table 1 and Figure 2, we note the following: (1) Our method generally outperforms the baseline model, demonstrating its effectiveness. (2) Furthermore, our method typically exceeds CVAR in most cases, highlighting the advantage of the diffusion method over the variational approach for cold-start CTR prediction tasks.

Ablation Study. We conduct ablation tests on our CSDM to determine the impact of its parameters on performance. First, we evaluate the performance of CSDM across different warm-up phases with varying ρ. The experiments are conducted over the TaobaoAD dataset using DeepFM as the backbone model. The results are presented in Table 2.

It is observable that the CSDM's performance in relation to ρ varies between the cold phase and the warm-up phase. In the cold phase, an increase in ρ generally leads to improved performance. Conversely, in the warm-up stage, an excessively large ρ results in a notable decrease in performance. The reason is that in the cold phase, the cold items

Dropout	Dataset	Cold	Warm-a	Warm-b	Warm-c
w	ML-1M	0.7456	0.7980	0.8059	**0.8091**
w/o	ML-1M	0.7456	0.7980	0.8059	0.8090
w	TaobaoAD	**0.6002**	0.6296	0.6324	**0.6383**
w/o	TaobaoAD	0.5859	**0.6306**	**0.6329**	0.6379
w	CIKM	**0.7418**	**0.7622**	**0.7687**	**0.7711**
w/o	CIKM	0.7402	0.7587	0.7677	0.7705

Table 3: An ablation test on the dropout function of diffusion models: "w" indicates that dropout is enabled, whereas "w/o" signifies that dropout is disabled. ML-1M stands for MovieLens-1M.

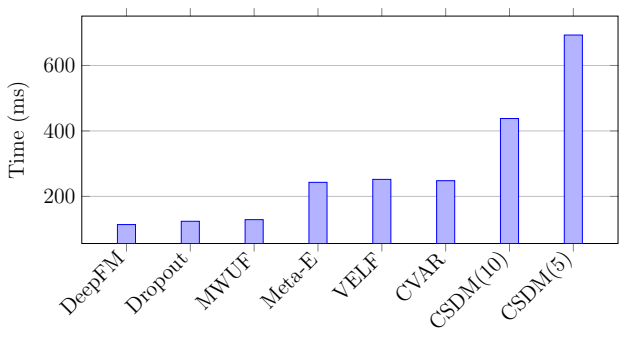

Figure 3: The time cost for training one batch using various methods, with CSDM tested using both $s = 5$ and $s = 10$.

have only the side information and $\mathcal{L}_{ctr}$ only contributes to the hot items. Increasing ρ can emphasize the contribution of the diffusion model in this setting. Additionally, this finding demonstrates the effectiveness of the CSDM approach in the cold-start problem. Meanwhile, in the warm-up stage, new items also have some user interaction. Enlarging ρ excessively can harm the contribution of user action data.

We conduct an ablation analysis on the diffusion model concerning the dropout mechanism across three datasets. The results are shown in Table 3. It can be observed that the effect of dropout varies across the datasets. On the MovieLens-1M dataset, the inclusion of dropout provides minimal improvement. In contrast, on the CIKM 2019 dataset, incorporating dropout into the diffusion process positively impacts and enhances the model's performance.

Complexity. Our method requires more forward passes than the CVAR method due to the diffusion process. The number of forward passes is proportional to the length of the sampled sequence. However, thanks to the non-Markovian characteristic, we can sample a sub-sequence to accelerate the generative process. We measure the time cost for training one batch on the MovieLens-1M dataset and report the results in Figure 3. The tests were conducted on a single A800 GPU with a batch size set to 2048. DeepFM is used as the backbone model. Since the warm-up embeddings are written back to the original ID embeddings, there is no additional computational overhead during inference.

Related Work

Cold-Start Recommendation. Many approaches have been proposed to improve the recommendation of new users or items. Some of these are model-dependent, for example, Heater (Zhu et al. 2020) and CLCRec (Wei et al. 2021) take CF-based models as their backbone. While some other methods are model-agnostic. For instance, DropoutNet (Volkovs, Yu, and Poutanen 2017) enhances the representation of users/items by applying dropout to exploit the average representations of users/items. MWUF (Zhu et al. 2021) introduces a Meta Scaling and Shifting Network to enhance the cold ID embeddings. Meta-E (Pan et al. 2019) recasts the CTR prediction task as a Meta-learning (Finn, Abbeel, and Levine 2017) problem and proposes a Meta-Embedding generator to initialize the cold ID embeddings. VELF (Xu et al. 2022) and CVAR (Zhao et al. 2022) learn probabilistic embeddings to alleviate the cold-start problem in CTR prediction. Similar to these methods, our method is also model-agnostic.

Diffusion Model in Recommendation. Since the success of DDPM (Ho, Jain, and Abbeel 2020) in image synthesis tasks, many researchers have attempted to leverage diffusion models in recommendation systems. DiffRec (Wang et al. 2023) employs diffusion models for generative recommendation. Diffusion models are also employed in (Ziqiang Cui 2024) for sequential recommendation, where they are used for the semantic generation of augmented views for contrastive learning (den Oord, Li, and Vinyals 2018). (Du et al. 2023) utilizes diffusion models to address the model collapse problems of variational auto-encoders in the sequential recommendation. (Wu et al. 2023) demonstrates that combining diffusion models with curriculum learning is beneficial for sequential recommendation.

Although some methods have explored the use of diffusion models for recommendations, to the best of our knowledge, employing diffusion models for cold-start in CTR prediction remains underdeveloped. This may be due to the challenges in cold-start scenarios, where we must construct a transition between ID embeddings and side information, whereas diffusion models are not directly applicable.

Conclusion

In this paper, we introduce a novel diffusion model to address the cold-start challenges in CTR prediction. It treats the embedding learning as a diffusion process. We design a non-Markovian diffusion process that enables the construction of an information flow between the side information of items and the pre-trained new item ID embeddings. Furthermore, our method can utilize both the collaborative filtering information from user action data and the side information in item features, making it applicable in both the cold-start and warm-up stages. Experiments conducted across three distinct recommendation datasets demonstrate that the proposed method is effective in addressing the cold-start problem in CTR prediction.

References

Chen, Z.; Xiao, R.; Li, C.; Ye, G.; Sun, H.; and Deng, H. 2020. ESAM: Discriminative Domain Adaptation with Non-Displayed Items to Improve Long-Tail Performance. In *Proceedings of the 43th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Cheng, H.-T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; Anil, R.; Haque, Z.; Hong, L.; Jain, V.; Liu, X.; and Shah, H. 2016. Wide & Deep Learning for Recommender Systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*.

Cui, Z.; Wu, H.; He, B.; Cheng, J.; and Ma, C. 2024. Diffusion-based Contrastive Learning for Sequential Recommendation. arXiv:2405.09369.

den Oord, A. V.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Du, H.; Yuan, H.; Huang, Z.; Zhao, P.; and Zhou, X. 2023. Sequential Recommendation with Diffusion Models. arXiv:2304.04541.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*.

Guo, H.; Chen, B.; Tang, R.; Zhang, W.; Li, Z.; and He, X. 2021. An Embedding Learning Framework for Numerical Features in CTR Prediction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*.

Lee, H.; Im, J.; Jang, S.; Cho, H.; and Chung, S. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Li, X. L.; Thickstun, J.; Gulrajani, I.; Liang, P.; and Hashimoto, T. B. 2022. Diffusion-LM improves controllable text generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Ling, C. X.; Huang, J.; and Zhang, H. 2003. AUC: a Statistically Consistent and more Discriminating Measure than Accuracy. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*.

Lu, Y.; Fang, Y.; and Shi, C. 2020. Meta-learning on Heterogeneous Information Networks for Cold-start Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Mo, K.; Liu, B.; Xiao, L.; Li, Y.; and Jiang, J. 2015. Image feature learning for cold start problem in display advertising. In *Proceedings of the 24th International Conference on Artificial Intelligence*.

Ouyang, W.; Zhang, X.; Ren, S.; Li, L.; Zhang, K.; Luo, J.; Liu, Z.; and Du, Y. 2021. Learning Graph Meta Embeddings for Cold-Start Ads in Click-Through Rate Prediction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Pan, F.; Li, S.; Ao, X.; Tang, P.; and He, Q. 2019. Warm Up Cold-start Advertisements: Improving CTR Predictions via Learning to Learn ID Embeddings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Rendle, S. 2010. Factorization Machines. *2010 IEEE International Conference on Data Mining*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597.

Roy, S.; and Guntuku, S. C. 2016. Latent Factor Representations for Cold-Start Video Recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*.

Saveski, M.; and Mantrach, A. 2014. Item cold-start recommendations: learning local collective embeddings. In *Proceedings of the 8th ACM Conference on Recommender Systems*.

Schein, A. I.; Popescul, A.; Ungar, L. H.; and Pennock, D. M. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Seroussi, Y.; Bohnert, F.; and Zukerman, I. 2011. Personalised rating prediction for new users using latent factor models. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*.

Shi, W.; Zhou, H.; Miao, N.; and Li, L. 2020. Dispersed exponential family mixture VAEs for interpretable text generation. In *Proceedings of the 37nd International Conference on International Conference on Machine Learning*.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. arXiv:2010.02502.

Tang, H.; Liu, J.; Zhao, M.; and Gong, X. 2020. Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*.

Teng, J.; Zheng, W.; Ding, M.; Hong, W.; Wangni, J.; Yang, Z.; and Tang, J. 2024. Relay Diffusion: Unifying diffusion process across resolutions for image synthesis. In *The Twelfth International Conference on Learning Representations*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Volkovs, M.; Yu, G.; and Poutanen, T. 2017. DropoutNet: addressing cold start in recommender systems. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Wang, R.; Fu, B.; Fu, G.; and Wang, M. 2017. Deep & Cross Network for Ad Click Predictions. In *Proceedings of the ADKDD'17*.

Wang, W.; Xu, Y.; Feng, F.; Lin, X.; He, X.; and Chua, T.-S. 2023. Diffusion Recommender Model. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Wei, Y.; Wang, X.; Li, Q.; Nie, L.; Li, Y.; Li, X.; and Chua, T.-S. 2021. Contrastive Learning for Cold-Start Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*.

Wu, Z.; Wang, X.; Chen, H.; Li, K.; Han, Y.; Sun, L.; and Zhu, W. 2023. Diff4Rec: Sequential Recommendation with Curriculum-scheduled Diffusion Augmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*.

Xu, X.; Yang, C.; Yu, Q.; Fang, Z.; Wang, J.; Fan, C.; He, Y.; Peng, C.; Lin, Z.; and Shao, J. 2022. Alleviating Cold-start Problem in CTR Prediction with A Variational Embedding Learning Framework. In *Proceedings of the ACM Web Conference 2022*.

Yan, L.; and Li, W.-J. 2014. Coupled group lasso for web-scale CTR prediction in display advertising. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*.

Zhang, J.; Zhao, C.; Ni, B.; Xu, M.; and Yang, X. 2019. Variational Few-Shot Learning. In *IEEE/CVF International Conference on Computer Vision*.

Zhao, X.; Ren, Y.; Du, Y.; Zhang, S.; and Wang, N. 2022. Improving Item Cold-start Recommendation via Model-agnostic Conditional Variational Autoencoder. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhou, G.; Mou, N.; Fan, Y.; Pi, Q.; Bian, W.; Zhou, C.; Zhu, X.; and Gai, K. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*.

Zhou, G.; Zhu, X.; Song, C.; Fan, Y.; Zhu, H.; Ma, X.; Yan, Y.; Jin, J.; Li, H.; and Gai, K. 2018. Deep Interest Network for Click-Through Rate Prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Zhu, Y.; Xie, R.; Zhuang, F.; Ge, K.; Sun, Y.; Zhang, X.; Lin, L.; and Cao, J. 2021. Learning to Warm Up Cold Item Embeddings for Cold-start Recommendation with Meta Scaling and Shifting Networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhu, Z.; Sefati, S.; Saadatpanah, P.; and Caverlee, J. 2020. Recommendation for New Users and New Items via Randomized Training and Mixture-of-Experts Transformation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ziqiang Cui, B. H. J. C. C. M., Haolun Wu. 2024. Diffusion-based Contrastive Learning for Sequential Recommendation. arXiv:2405.09369.