Modeling Long-term User Behaviors with Diffusion-driven Multi-interest Network for CTR Prediction

Weijiang Lai
Institute of Software, Chinese
Academy of Sciences
University of Chinese Academy of
Sciences, Beijing, China
laiweijiang22@otcaix.iscas.ac.cn

Beihong Jin[†]
Institute of Software, Chinese
Academy of Sciences
University of Chinese Academy of
Sciences, Beijing, China
Beihong@iscas.ac.cn

Yapeng Zhang
Meituan
Beijing, China
zhangyapeng05@meituan.com

Yiyuan Zheng
Institute of Software, Chinese
Academy of Sciences
University of Chinese Academy of
Sciences, Beijing, China
zhengyiyuan22@otcaix.iscas.ac.cn

Rui Zhao
Institute of Software, Chinese
Academy of Sciences
University of Chinese Academy of
Sciences, Beijing, China
zhaorui22@otcaix.iscas.ac.cn

Jian Dong
Meituan
Beijing, China
dongjian03@meituan.com

Jun Lei
Meituan
Beijing, China
leijun@meituan.com

Xingxing Wang
Meituan
Beijing, China
wangxingxing04@meituan.com

ABSTRACT

CTR (Click-Through Rate) prediction, crucial for recommender systems and online advertising, etc., has been confirmed to benefit from modeling long-term user behaviors. Nonetheless, the vast number of behaviors and complexity of noise interference pose challenges to prediction efficiency and effectiveness. Recent solutions have evolved from single-stage models to two-stage models. However, current two-stage models often filter out significant information, resulting in an inability to capture diverse user interests and build the complete latent space of user interests. Inspired by multi-interest and generative modeling, we propose DiffuMIN (Diffusion-driven Multi-Interest Network) to model long-term user behaviors and thoroughly explore the user interest space. Specifically, we propose a target-oriented multi-interest extraction method that begins by orthogonally decomposing the target to obtain interest channels. This is followed by modeling the relationships between interest channels and user behaviors to disentangle and extract multiple user interests. We then adopt a diffusion module guided by contextual interests and interest channels, which anchor users' personalized and target-oriented interest types, enabling the generation of augmented interests that align with the latent spaces of user interests, thereby further exploring restricted interest space. Finally, we leverage contrastive learning to ensure that the generated augmented interests align with users' genuine preferences. Extensive offline experiments are conducted on two public datasets and one industrial dataset, yielding results that demonstrate the superiority of DiffuMIN. Moreover, DiffuMIN increased CTR by 1.52% and CPM by 1.10% in online A/B testing. Our source code is available at https://github.com/laiweijiang/DiffuMIN.

CCS CONCEPTS

• **Information systems → Recommender systems**; **Online advertising**; **Learning to rank**.

KEYWORDS

CTR Prediction, User Behavior Modeling, Diffusion Model

[†] Corresponding author.

1 INTRODUCTION

CTR (Click-Through Rate) prediction, which estimates the probability of a user clicking a target item, has been a hot topic for both academia and industry. Recent studies and practices show that modeling long-term user behavior can improve CTR performance.

Modeling long-term behaviors faces difficulties in effectiveness and efficiency. First, it suffers from coupled interests and noise interference, hindering accurate interest extraction. Second, the lengthy behavior sequences make it infeasible to employ mechanisms with high time complexity, such as self-attention [54].

To address these challenges, current solutions have evolved from single-stage models to two-stage models [2–4, 60]. Early single-stage models focus on optimizing the time complexity of target-attention or self-attention mechanisms for comprehensive modeling of long-term user behaviors. However, these simplified attention mechanisms often fail to capture user interests accurately and comprehensively. In contrast, two-stage models first filter behaviors and obtain ones that are more relevant to the target using similarity scores and then model these behaviors separately. These solutions alleviate both efficiency and noise interference issues, making them prevalent. However, despite recent advancements in two-stage approaches [15, 25, 31, 47], such as incorporating richer features to aid in behavior filtering, these models tend to extract behaviors from a single perspective, resulting in homogeneity and redundancy that significantly constrain the user's interest space and ultimately limit their performance.

Compared to short-term user behaviors, long-term behaviors encompass abundant user interests. Inspired by multi-interest modeling, we propose filtering user behaviors from multiple perspectives within long-term behaviors to disentangle and extract multiple user interests. Further, although we maximize the preservation of multiple user interests through multi-interest modeling, the filtering mechanism inherently restricts the user's interest space. Inspired by generative modeling, we incorporate a diffusion module to learn the distribution of users' multiple interests, capturing nuanced and robust augmented interests. This process expands the original interests within the distribution, thereby revealing new insight for exploration and understanding user interests.

To explore user interest space and unleash the potential of long-term behaviors, we propose DiffuMIN (Diffusion-driven Multi-Interest Network), a two-stage model for long-term behavior modeling. In the first stage, we propose a target-oriented multi-interest extraction method that begins by decomposing the target embedding into orthogonal interest channels and modeling the relationship between user behaviors and these channels. It extracts aggregated interests by routing each behavior to the channel with the highest score in relevance and aggregating behaviors with scores in the top-$p\%$ in each channel, effectively reducing inter-channel redundancy and filtering out irrelevant behaviors in long-term sequences. In the second stage, we design a diffusion module to generate multiple augmented interests to supplement aggregated interests. To the best of our knowledge, we are the first to apply diffusion modeling to user interests in CTR prediction. This diffusion process is guided by contextual interests and interest channels, which anchor users' personalized and target-oriented interest types, respectively. Additionally, instead of conventional Gaussian noise, we employ perturbed user interests as the starting points for initial generation, thereby enhancing personalization and simplifying the sampling process for high-quality representations. Lastly, to further optimize generation quality, we introduce contrastive learning to ensure the augmented interests align with a user's actual preferences, enhancing the distinguishability of interests among different users.

Our contributions are summarized as follows.

- We propose the target-oriented multi-interest extraction method, which disentangles and extracts multiple user interests by modeling the relationship between user behaviors

and interest channels derived from the orthogonal decomposition of the target, thereby deriving the diverse user interests.
- We propose the diffusion module guided by contextual interests and interest channels, enabling the generation of augmented interests that align with the latent spaces of user interests, thereby sustaining and enriching the latent interest space.
- We conduct extensive offline experiments on three real-world datasets and online A/B testing. Experimental results show that DiffuMIN achieves SOTA performance.

2 RELATED WORK

Our research primarily involves CTR models for long-term behaviors, multi-interest modeling, and generative modeling.

2.1 CTR Models for Long-term Behaviors

CTR models predict the likelihood of interactions between a user and a target. One prominent category focuses on modeling feature interactions, such as Wide&Deep and DeepFM [6–10, 34, 35, 37]. Additionally, another category emphasizes analyzing user behaviors to boost CTR predictions[5, 12, 13]. The models that fall into this category include DIN [1], which utilizes an attention mechanism to prioritize relevant behaviors, DIEN [11], which captures evolving interests with attention and GRU, and HSTU [53], which integrates recall and ranking tasks within a unified architecture, exploring scaling laws in recommendation scenarios.

Aside from these models, numerous studies focus on modeling long-term user behaviors to uncover behavior dependencies and periodic patterns within user behaviors. Early models like MIMN [14] and HPMN [16] utilize memory networks to manage user interests but encounter challenges in providing timely updates and modeling the relationships between behaviors and targets, which ultimately limits performance.

Recent models such as SDIM[2] and LinRec [60] attempt to simplify attention mechanisms for fully long-term behavior modeling. However, these simplified approaches often struggle to effectively model the nuances of long-term behavior relationships and capture complex user behavior patterns. In contrast, other models, including SIM [3], begin by identifying the top-k behaviors most relevant to the target item, then modeling these behaviors, respectively. Building on this two-stage line of thought, models such as UBR4CTR [15], TWIN [31], and CoFARS [47] incorporate auxiliary information to improve the accuracy of filtering in the first stage [15, 31]. Meanwhile, ETA [4] and TWIN-v2 [25] employ proven techniques such as SimHash [17] encoding and clustering to boost model efficiency.

Although long-term behaviors offer a wealth of user interests, current methods, particularly two-stage models, despite achieving certain results, inevitably constrain the expression of user interests, thereby limiting overall performance.

2.2 Multi-interest Modeling Methods

Multi-interest modeling, which enhances the performance by identifying users' diverse and dynamically changing interests, has become a focal area of research [49–51, 55, 56].

For instance, MIND [22] and ComiRec [59] utilize dynamic routing and capsule networks to adaptively aggregate user behaviors into multiple embeddings, representing multiple user interests. DMIN [57] employs multi-head self-attention to encode user behaviors, modeling the output of each head with the target as distinct user interests. Octopus [48] initializes multiple orthogonal interest channels to aggregate user behaviors for multi-interest extraction. Trinity [58] adopts a two-clustering approach, extracting multiple user interests from the primary cluster and secondary cluster.

However, these approaches either concentrate solely on modeling relationships within user behaviors to extract multiple interests or use high-complexity techniques like self-attention, making them unsuitable for modeling long-term user behaviors in CTR scenarios.

2.3 Generative Modeling Methods

Common generative models include Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models [38, 39, 70, 71], with diffusion models demonstrating superior theoretical and practical performance, achieving state-of-the-art results in fields such as image generation [33, 36]. Inspired by this, researchers try to introduce diffusion models to the recommender systems, aiming to better capture complex distributions of user behavior and features, while alleviating data sparsity issues [40–46].

For example, DreamRec [67] employs guided diffusion to generate oracle items aligned with user interests, recommending real items that best match these oracle items. DiffRec [66] modifies the conventional sampling starting point from Gaussian noise to perturbed embeddings, reducing noise within the original embeddings. DiffuASR [65] uses diffusion models to learn the distribution of user behavior embeddings, directly generating sequences to augment behavior data. PDRec [64] and Diff4Rec [63] apply similar diffusion modules as DiffRec for augmentation. CaDiRec [62] also implements diffusion to generate data guided by positional encoding and context, employing these for contrastive learning. SeeDRec [61] introduces sememes as a generation granularity, enhancing existing models with additional information to boost performance.

However, diffusion techniques have not been found to be applied to CTR prediction for modeling user interests.

3 METHODOLOGY

3.1 Problem Formulation

For each user u, which has a user behavior sequence $\mathcal{S} = \{b_i\}_{i=1}^{l}$ and more detailed features such as age and gender. Here, b_i is the i-th behavior and l is the length of the user behavior sequence. Each behavior b_i has features like item ID and category ID. Regarding a target item s, it includes features such as item ID and more additional contextual features.

Given a user u and a target item s, with their interaction label $y \in \{0, 1\}$, our task is to predict the click-through probability, formalized as follows:

$$\mathcal{P}(y = 1|x \in \{(u, s)\}) = F(x; \theta), \tag{1}$$

where $F(x; \theta)$ is the model we will develop and θ represents the parameters of the model.

For this task, we propose a model named DiffuMIN. Figure 1 shows the architecture of our model, which mainly includes the input layer, Orthogonal Multi-Interest Extractor (OMIE), Diffusion Multi-Interest Generator (DMIG), Contrastive Multi-Interest Calibrator (CMIC), and prediction layer.

3.2 Input Layer

In this layer, we utilize a uniform embedding table to initialize all embeddings, including the user behavior sequence embedding matrix $E = [e_1, ..., e_l] \in \mathbb{R}^{l \times d}$, where e_i represents the embedding of the i-th behavior, the target item embedding vector $e_s \in \mathbb{R}^d$, and embedding vector e_{other} representing other features. The embedding dimension d is used uniformly to denote the dimension after the concatenation of features.

In this layer, we extract the recent k behaviors to form the user short-term behavior embedding matrix $E_k = [e_1, \ldots, e_k] \in \mathbb{R}^{k \times d}$, which is input into the Transformer encoder with the target embedding e_s to model the user's current interest. Simultaneously, we feed the complete behavior embedding matrix E and the target embedding e_s into the OMIE to disentangle and extract multiple user interests.

3.3 Orthogonal Multi-Interest Extractor

Compared to short-term behaviors, long-term user behaviors contain a wider range of interests. However, capturing these interests comprehensively poses a challenge due to coupled interests and numerous irrelevant behaviors within long-term behaviors. Existing two-stage models often filter behaviors based solely on their similarity to targets, leading to repetitive behaviors and interest redundancy, which substantially restrict the user's interest space.

Inspired by multi-interest modeling, we propose OMIE, designed to disentangle and extract user interests from various perspectives within long-term behaviors, mitigating the reduction of interests common in two-stage approaches. In this module, we introduce a target-oriented multi-interest extraction method utilizing the orthogonal decomposition of target embeddings as interest channels. By learning relationships between behaviors and these diverse interest channels, we decouple and extract multiple user interests, thus preserving the user's interest space. Unlike traditional multi-interest models that typically rely on neural networks to explore intra-behavior relationships for multi-interest extraction, our approach emphasizes the relationship between behaviors and the target, which is vital in CTR scenarios.

Specifically, we first perform a linear projection to the target item embedding, followed by orthogonal decomposition to derive a set of basis embeddings as interest channels. The specific formula is as follows:

$$e_s' = e_s W, \tag{2}$$

$$O \leftarrow \arg\min_{O} \sum_{e_s'} \left\| e_s' - O^T O e_s' \right\|_2, \tag{3}$$

subject to

$$O^T O = \mathbf{I}, \tag{4}$$

where $W \in \mathbb{R}^{d \times cd}$, $O = [o_1, o_2, \ldots, o_c] \in \mathbb{R}^{c \times d}$ and c is the number of interest channels. Each o_i represents an interest channel, capturing a specific target-oriented aspect of user interest. The
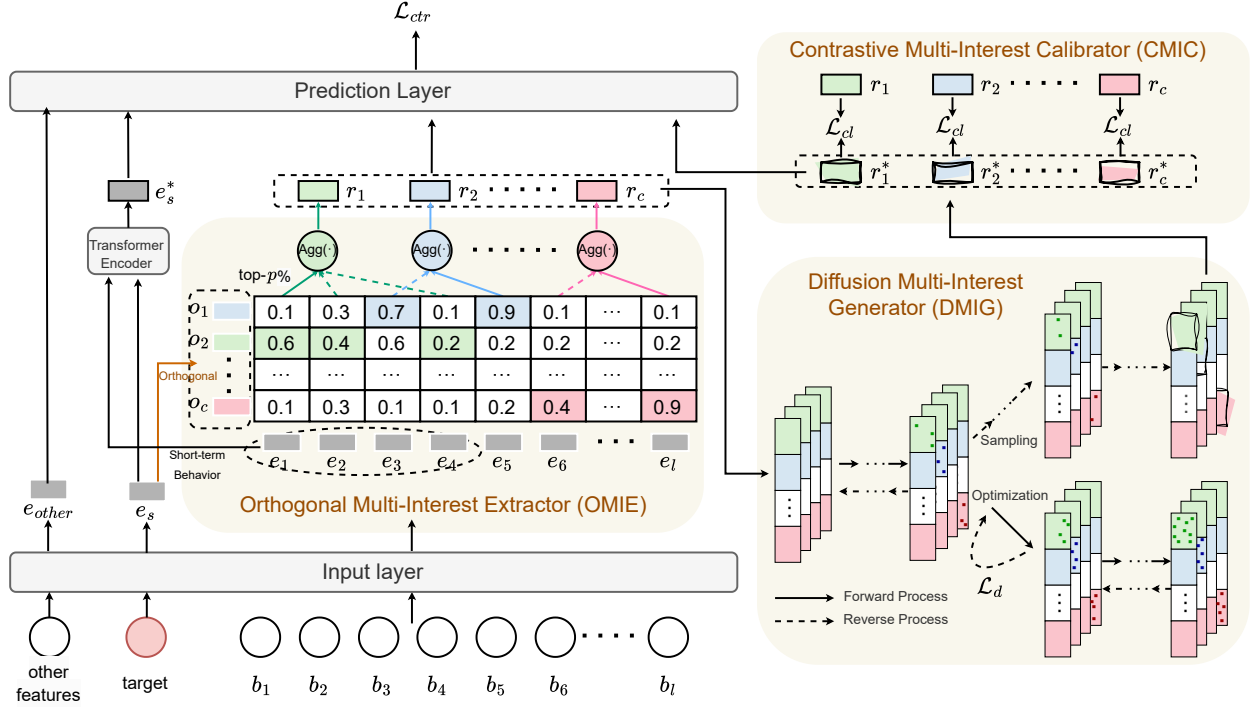
Figure 1: Architecture of DiffuMIN.

orthogonality of these channels efficiently reduces interest redundancy in long-term behaviors and enhances the effectiveness of multi-interest extraction.

We determine the relationship between user behaviors and interest channels through matrix multiplication of the user behavior matrix E with the interest channel matrix O, expressed as:

$$A = EO^T. \tag{5}$$

In matrix $A \in \mathbb{R}^{l \times c}$, each element $A_{i,j}$ represents the score indicating the relevance between behavior e_i and channel o_j.

Within the long-term behavior scenario, we accurately extract multiple user interests through a process involving behavior routing, channel filtering, and interest aggregation.

Behavior Routing. This step identifies the most relevant interest channel for each user behavior. We apply a top-1 routing approach, ensuring each behavior is routed exclusively to the channel where it scores the highest in relevance. This minimizes redundancy across channels, enhancing specificity and match accuracy.

$$\Phi_{i,j} = \begin{cases} 1, & \text{if } j = \arg\max_{j'} A_{i,j'} \\ 0, & \text{otherwise} \end{cases} . \tag{6}$$

Channel Filtering. This step aggregates only genuinely relevant behaviors within each interest channel. We filter irrelative behaviors and retain the top-$p\%$ of behaviors for each channel based on their scores, thereby ensuring only the most pertinent behaviors are kept in long-term sequences, thus maintaining the accuracy of

user interests.

$$\Gamma_{i,j} = \begin{cases} 1, & \text{if } \Phi_{i,j} = 1 \text{ and } A_{i,j} \in \text{top-}p\% \text{ of } A_{:,j} \\ 0, & \text{otherwise} \end{cases} . \tag{7}$$

Interest Aggregation. In this step, we aggregate the remaining behaviors from each interest channel to form an aggregated interest.

$$r_j = \text{Agg} \left(\{ e_i \mid \Gamma_{i,j} = 1 \} \right), \tag{8}$$

where we utilize mean pooling as the aggregation function.

3.4 Diffusion Multi-Interest Generator

While two-stage models effectively reduce irrelevant behaviors in long-term sequences, they concentrate on a limited subset of user behaviors to extract interests, inevitably missing out on subtle and latent user interests. Despite employing OMIE to preserve a broad range of user interests, it still inherently limits the latent space of user interests.

Inspired by generative modeling, we introduce a diffusion module to capture the distribution of multiple user interests and generate augmented interest that conforms to the latent spaces of user interests, complementing the original aggregated interest.

In traditional diffusion models, a parameterized Markov chain gradually corrupts source data with controlled noise in the forward process, transforming it into Gaussian noise. The reverse process reconstructs the data step-by-step from Gaussian noise, optimizing the network in the diffusion phase and generating data in the sampling phase.

Building on these concepts, we propose an enhanced diffusion module guided by contextual interests and interest channels. This module leverages conditional information to anchor personalized interest scopes and types effectively. Rather than starting from Gaussian noise during the sampling phase, we begin with perturbed user interests. This approach enhances personalization and streamlines the sampling process, facilitating the generation of high-quality representations.

3.4.1 Diffusion Optimization Phase. During the diffusion optimization phase, the diffusion network learns the distribution of each user across various interest types. In the forward process, the time step t is uniformly sampled from $\{1, 2, \ldots, T\}$ to introduce noise into aggregated interest r_i, where $i \in [1, c]$. A key characteristic of the forward process is its ability to directly sample the perturbed result $r_{i,t}$ of r_i at time step t given schedule β_t, as follows:

$$q\left(r_{i,t} \mid r_{i,0}\right) = \mathcal{N}\left(r_{i,t}; \sqrt{\bar{\alpha}_t}r_{i,0}, (1 - \bar{\alpha})\mathbf{I}\right), \tag{9}$$

$$r_{i,t} = \sqrt{\bar{\alpha}_t}r_{i,0} + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{10}$$

where $r_{i,0} = r_i$, $\epsilon \sim \mathcal{N}(0, 1)$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{j=1}^{t} \alpha_j$, $T = 1000$ and β_t follows a linear variance schedule from 0.0001 to 0.02.

In the reverse process, the reconstruction of perturbed interest $r_{i,t}$ is guided by two conditions: the contextual interest condition $g_1 = [r_1, \ldots, r_{i-1}, r_{i+1}, \ldots, r_c]$ and the interest channel condition $g_2 = o_i$. The reverse process is formulated as follows:

$$p_\theta(r_{i,t-1}|r_{i,t},g_1,g_2) = \mathcal{N}(r_{i,t-1};\mu_\theta(r_{i,t},t,g_1,g_2), \Sigma_\theta(r_{i,t},t,g_1,g_2)), \tag{11}$$

In the implementation, we only optimize the mean [38], and reconstruct $r_{i,t-1}$ from $r_{i,t}$ as follows:

$$\mu_\theta(r_{i,t},t,g_1,g_2) = \frac{1}{\sqrt{\alpha_t}}(r_{i,t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(r_{i,t},t,g_1,g_2)), \tag{12}$$

$$r_{i,t-1} = \frac{1}{\sqrt{\alpha_t}}(r_{i,t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(r_{i,t},t,g_1,g_2)) + \sigma_t z, \tag{13}$$

where $z \sim \mathcal{N}(0, 1)$, and $\sigma_t = \sqrt{\beta_t}$ is the standard deviation. In particular, we use the Transformer encoder as the backbone for the diffusion network ϵ_θ to predict the added noise at step t when generating $r_{i,t-1}$ from $r_{i,t}$. The diffusion network primarily incorporates self-attention and cross-attention mechanisms to support the guidance of contextual interests and interest channels.

Contextual Interests Guidance. This component anchors the user's personalized interests by establishing the relationship between $r_{i,t}$ and contextual interest g_1 through the self-attention layer. Specifically, $r_{i,t}$ is concatenated with g_1 to form the matrix $R_{i,t} = [r_1, \ldots, r_{i-1}, r_{i,t}, r_{i+1}, \ldots, r_c]$, which is then processed by the self-attention layer as follows:

$$Q = R_{i,t}W_{self}^Q, \; K = R_{i,t}W_{self}^K, \; V = R_{i,t}W_{self}^V, \tag{14}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{15}$$

where $W_{self}^Q, W_{self}^K, W_{self}^V \in \mathbb{R}^{d \times d}$ are learnable projection matrices.

Interest Channels Guidance. This component anchors target-oriented interest types by analyzing the relationship between user

---

Algorithm 1 Diffusion Optimization Phase

Input: $r_1, \ldots, r_c, o_1, \ldots, o_c$
Output: $\mathcal{L}_d$
  $\mathcal{L}_d = 0$
  for $i \leftarrow 1$ to c **do**
    $r_{i,0} \sim q(r_i)$
    $t \sim \text{Uniform}(\{1, \ldots, T\})$
    $\epsilon \sim N(0, 1)$
    $g_1 = [r_1, \ldots, r_{i-1}, r_{i+1}, \ldots, r_c], \; g_2 = o_i$
    $\mathcal{L}_d += \left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t}r_{i,0} + \sqrt{1 - \bar{\alpha}_t}\epsilon, t, g_1, g_2 \right) \right\|_2^2$
  end for

---

interest $r_{i,t}$ and interest channel g_2 using cross-attention. The specific formulas are as follows:

$$Q' = R_{i,t}W_{cross}^Q, \; K' = g_2W_{cross}^K, \; V' = g_2W_{cross}^V, \tag{16}$$

$$\text{Attention}(Q', K', V') = \text{softmax}\left(\frac{Q'K'^T}{\sqrt{d}}\right)V', \tag{17}$$

where $W_{cross}^Q, W_{cross}^K, W_{cross}^V \in \mathbb{R}^{d \times d}$ are learnable projection matrices. In this phase, $r_{i,t} \in R_{i,t}$ is an intermediate representation, while other embeddings g_1 and g_2 are fixed. The optimization loss for the diffusion module is calculated as follows:

$$\mathcal{L}_d = \frac{1}{c}\sum_{i=1}^{c}\mathbb{E}_{r_{i,t},g_1,g_2,\epsilon \sim \mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(r_{i,t},t,g_1,g_2)||_2^2]. \tag{18}$$

The detailed steps of the diffusion optimization phase are outlined in Algorithm 1.

3.4.2 Diffusion Sampling Phase. Our diffusion sampling phase diverges from traditional models that sample directly from Gaussian noise. For the i-th interest, we perform the forward process to obtain a perturbed interest $r_{i,t}$ as the starting point for sampling. This approach preserves user personalization while effectively removing noise from r_i, simplifying diffusion generation. Consequently, the reverse process, which iteratively reconstructs $r_{i,t-1}^*$ over $T'(T' << T)$ steps, is sufficient to progressively denoise, yielding the optimized interest representation as the user's augmented interest at $t = 0$, i.e., $r_i^* = r_{i,0}^*$. The formula is as follows:

$$r_{i,t-1}^* = \frac{1}{\sqrt{\alpha_t}}\left(r_{i,t}^* - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(r_{i,t}^*,t,g_1,g_2)\right) + \sigma_t z, \; 0 < t \leq T'. \tag{19}$$

During the diffusion sampling phase, we generate c different perspectives of augmented interests $r^* = [r_1^*, \ldots, r_c^*]$ for each user, thereby enriching the corresponding aggregated interests. The specifics of the diffusion sampling phase are detailed in Algorithm 2.

To achieve end-to-end training and directly enhance CTR performance through multiple augmented interests, we perform both the diffusion optimization and sampling phases during model training, while only the diffusion sampling phase is done during inference.

3.5 Contrastive Multi-Interest Calibrator

In OMIE and DMIG, we obtain multiple aggregated interests $r = [r_1, \ldots, r_c]$ through important behavior aggregation, and augmented interests $r^* = [r_1^*, \ldots, r_c^*]$ via diffusion sampling. To further explore

Algorithm 2 Diffusion Sampling Phase

Input: $r_1, ..., r_c, o_1, ..., o_c$
Output: $r_1^*, ..., r_c^*$
  for $i \leftarrow 1$ to c **do**
    $t \sim \text{Uniform}(\{1, ..., T\})$
    $r_{i,t} = \sqrt{\bar{\alpha}_t} r_{i,0} + \sqrt{1 - \bar{\alpha}_t} \epsilon$
    $r_{i,T'}^* \sim q(r_{i,t})$
    $g_1 = [r_1, ..., r_{i-1}, r_{i+1}, ..., r_c], \quad g_2 = o_i$
    for $t \leftarrow T'$ to 1 **do**
      $t > 1 ? z \sim \mathcal{N}(0, 1) : z = 0$
      $r_{i,t-1}^* = \frac{1}{\sqrt{\alpha_t}} (r_{i,t}^* - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(r_{i,t}^*, t, g_1, g_2)) + \sigma_t z$
    end for
    $r_i^* = r_{i,0}^*$
  end for

---

the relationships between these aggregated and augmented interests, we propose employing contrastive learning to examine their similarities and differences, thereby enhancing representational learning.

Contrastive learning is a self-supervised method that achieves superior representation learning by minimizing the distance between positive samples while maximizing the separation between negative samples, thereby improving the model's generalization and robustness. In this module, for a user's i-th aggregated interest r_i, the positive sample is the user's augmented interest r_i^*, while negative samples are the augmented interest r_i^* of other users within the mini-batch. In particular, We project these embeddings to other spaces and optimize them using the following loss function to prevent interference with the primary CTR task:

$$\mathcal{L}_{cl} = \frac{1}{c|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{i=1}^{c} -\log \frac{\exp\left(\text{sim}(r_i^{(u)}, r_i^{*(u)})/\tau\right)}{\sum_{v \in \mathcal{U}} \exp\left(\text{sim}(r_i^{(u)}, r_i^{*(v)})/\tau\right)}, \quad (20)$$

where $\mathcal{U}$ denotes the set of users in the mini-batch and τ is the temperature. This loss function not only aligns the augmented interests with the user's actual preferences but also enhances the distinguishability of interests among different users.

3.6 Prediction Layer

In this layer, we input the aggregated interests r, augmented interests r^*, the user's short-term interest e_s^* encoded by the Transformer encoder, and e_{other} into the multilayer perceptron (MLP) for CTR prediction, as described by the following formula:

$$e_s^* = \text{Transformer}([e_1, ..., e_k, e_s])[-1, :], \quad (21)$$

$$\hat{y} = \sigma(\text{MLP}([r, r^*, e_{other}, e_s^*])), \quad (22)$$

where $e_1, ..., e_k$ represent the embeddings of the user's short-term behaviors, e_s^* is the last element of Transformer encoder output, and σ is the sigmoid function. We use the binary cross-entropy loss to optimize our model, as follows:

$$\mathcal{L}_{ctr} = -\frac{1}{N} \sum_{i=1}^{N} (y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)), \quad (23)$$

where N is the number of samples, and y_i is the label of the i-th sample. During training, the three loss functions are optimized

Table 1: Statistics of the datasets.

Datasets	#Users	#Items	#Samples
Ele.me	14,427,689	7,446,116	128,000,000
Alibaba	1,141,729	461,527	700,000,000
Industry	84,262,000	16,624,521	1,061,530,7879

simultaneously as follows:

$$\mathcal{L} = \mathcal{L}_{ctr} + \lambda_1 \mathcal{L}_d + \lambda_2 \mathcal{L}_{cl}, \quad (24)$$

where λ_1 and λ_2 are the weighting coefficients.

3.7 Complexity Analysis

3.7.1 Space Complexity. In DiffuMIN, the additional learnable parameters primarily arise from the projection layers in OMIE and CMIC, and the diffusion network in DMIG. These have spatial complexities of $O(cd^2 + 2d^2)$ and $L(8d^2 + 2dd_f)$, where L denotes the number of diffusion network layers and d_f is the dimension of the feedforward layer in the diffusion network, respectively. Consequently, the parameter increase introduced by DiffuMIN is minimal.

3.7.2 Time Complexity. The time complexity of our model is comprised primarily of OMIE, DMIG, and CMIC, with respective complexities of $O(Blcd + Blc^2)$, $O(BT'L(cd^2 + c^2d + cdd_f))$, and $O(bcd^2 + bc^2d)$. Given that $c << l$ and T' is small in our configuration, the overall time complexity is approximately $O(blcd)$. Thus, our model's efficiency is comparable to recent models designed for long-term user behaviors.

4 EXPERIMENTS

In this section, we conduct extensive experiments to answer the following Research Questions (RQs):

- **RQ1**: Does DiffuMIN outperform existing CTR models when modeling long-term user behaviors?
- **RQ2**: What contributions do the individual modules of DiffuMIN make to its overall performance?
- **RQ3**: How do different designs within the OMIE module affect model performance?
- **RQ4**: What is the performance impact of key designs in the DMIG module?
- **RQ5**: How does DiffuMIN's interest modeling differ from traditional models in representing user interest spaces?
- **RQ6**: How does DiffuMIN perform in the live production environment?

4.1 Experimental Settings

4.1.1 Datasets. We select two public datasets and one industrial dataset to conduct experiments. The statistics of three datasets are shown in Table 1.

- **Alibaba**[1]: This dataset, provided by Alibaba, is a display advertising click-through rate prediction dataset. It includes shopping behavior data from all users over a 22-day period and includes comprehensive information on users, advertisements, and user behaviors.

---

[1]https://tianchi.aliyun.com/dataset/56

Table 2: Performance comparison. We conduct each experiment three times and report the average results. In each row, the best and second-best results are highlighted in bold and underlined, respectively. DIN(S) is considered the base model for calculating the RelaImpr.

Dataset	Metric	DIN(S)	CAN(S)	DIN	CAN	SoftSIM	HardSIM	ETA	SDIM	TWIN	TWIN-V2	DiffuMIN
Industry	AUC	0.6740	0.6736	0.6751	0.6748	0.6772	0.6780	0.6778	0.6779	0.6785	<u>0.6788</u>	**0.6841**
	RelaImpr	0.00%	-0.23%	0.63%	0.45%	1.84%	2.30%	2.18%	2.24%	2.59%	<u>2.76%</u>	**5.80%**
Alibaba	AUC	0.6125	0.6091	0.6198	0.6184	0.6212	<u>0.6239</u>	0.6220	0.6206	0.6215	0.6220	**0.6282**
	RelaImpr	0.00%	-3.02%	6.49%	5.24%	7.73%	<u>10.13%</u>	8.44%	7.20%	8.00%	8.44%	**13.96%**
Ele.me	AUC	0.6363	0.6378	0.6273	0.6284	0.6399	0.6389	0.6398	0.6404	<u>0.6410</u>	0.6400	**0.6462**
	RelaImpr	0.00%	1.10 %	-6.60%	-5.80%	2.64%	1.90%	2.57%	3.01%	<u>3.45%</u>	2.71%	**7.26%**

- **Ele.me**[2]: This dataset is constructed by click logs from ele.me online recommender system and contains 30-day behaviors of users. It includes features related to users, candidate items, user behaviors, and spatiotemporal features.
- **Industry**: It is an industrial dataset, that contains over 84 million users who have been active within the last 7 days and collects their complete behavior records for the past year, where each user behavior includes features like item ID, behavior type and so on.

4.1.2 Competitors. Our competitors include **DIN**[1], **CAN**[5], **SIM** [3] with its two versions **SoftSIM** and **HardSIM**, along with **ETA** [4], **SDIM** [2], **TWIN** [31] and **TWIN-V2** [25].

4.1.3 Evaluation Metrics. We adopt the widely used Area Under Curve (AUC) as the offline evaluation metric. For online experiments, we adopt CTR (Click-Through Rate), and CPM (Cost Per Mille) as evaluation metrics. Besides, we follow [1, 23] to introduce the relative improvement (Relalmpr) metric to measure relative improvement between models, defined as follows:

$$\text{RelaImpr} = \left(\frac{\text{AUC(model)} - 0.5}{\text{AUC(base model)} - 0.5} - 1 \right) \times 100\%. \quad (25)$$

4.1.4 Implementation Details. We implement all models by TensorFlow. For model training, we use Adam as the optimizer, and each model is trained for one epoch. Each feature dimension is set to 8. All models have the same configuration for fairness.

In DiffuMIN, we meticulously tune hyperparameters including the number of channels c within $\{2, 4, 8\}$, diffusion sampling step T' within $\{5, 10, 20, 50\}$, the temperature τ within $\{0.001, 0.005, 0.01, 0.05\}$, and the weights of auxiliary losses λ_1 and λ_2 within $\{0.0001, 0.001, 0.01, 0.1\}$. Our model achieves optimal performance on the industrial dataset with c, T', τ, λ_1, and λ_2 set to 4, 20, 0.05, 0.01, and 0.001, respectively.

On the industrial, Alibaba, and Ele.me datasets, due to differences in data duration, we set the maximum behavior sequence length l to 5000, 1500, and 50, padding shorter sequences with zeros. In DiffuMIN, each channel aggregates the top 20% of behaviors. The DIN and CAN models use short behavior inputs with lengths of 100, 100, and 20, resulting in DIN(S) and CAN(S). Meanwhile, two-stage models for long-term behaviors retain sequences of the second stage with lengths of 100, 100, and 20, respectively.

[2]https://tianchi.aliyun.com/dataset/131047

4.2 Overall Performance (RQ1)

Table 2 shows the performance results of competitors and DiffuMIN across three datasets. Note that an AUC improvement of 0.001 level is considered significant in CTR prediction scenarios [21, 31]. We also conduct a *t*-test on AUCs with a significance level of 0.05, indicating that there is a significant difference between DiffuMIN and comparisons in terms of performance.

The results of DIN and CAN reveal that merely extending the length of user behaviors in regular CTR models does not necessarily improve performance. More behaviors may provide more information but can overshadow crucial ones, so it is more advantageous to only model recent behaviors.

Models designed for long-term behaviors show superior ability in modeling long-term user behaviors compared to regular CTR models. Two-stage models are particularly effective, as they identify and retrieve crucial behaviors in the first stage and model them separately in the second stage. These models reduce model time complexity and mitigate irrelevant behaviors in long-term behaviors, significantly boosting performance.

DiffuMIN achieves optimal performance, highlighting the importance of extensively exploring user interests. We first derive multiple aggregated interests from various perspectives, followed by generating multiple augmented interests using a diffusion module, which substantially boosts performance.

4.3 Ablation Study (RQ2)

We conduct in-depth ablation experiments to analyze the contribution of each module within DiffuMIN.

Variants A and B examine the OMIE module from different perspectives. Variant A evaluates OMIE by filtering behaviors solely based on the similarity between target and behavior embeddings, then aggregating user interest and feeding it to subsequent modules. Variant B removes the multiple aggregated interests provided by OMIE in the prediction layer. Variant C omits CMIC, while Variant D removes both DMIG and CMIC. Lastly, variant E eliminates the module for modeling users' short-term behaviors.

The experimental results are presented in Table 3. Variant A's results demonstrate that using the target-oriented multi-interest extraction method to disentangle and extract multiple user interests effectively enhances model performance. Variant B's AUCs indicate that explicitly capturing multiple aggregated interests are essential, with augmented interests complementing the original ones.

Table 3: Results of the ablation study.

Models	Industry	Alibaba	Ele.me
DiffuMIN	**0.6841**	**0.6282**	**0.6462**
Variant A	0.6810	0.6257	0.6433
Variant B	0.6769	0.6221	0.6385
Variant C	0.6834	0.6274	0.6450
Variant D	0.6826	0.6260	0.6448
Variant E	0.6812	0.6250	0.6450

The results of variants C and D show that our diffusion module successfully generates multiple augmented interests, thoroughly exploring users' limited interests. Furthermore, the incorporation of contrastive learning effectively boosts the expressiveness of representations. Variant E's performance highlights the importance of extracting short-term interests, consistent with findings from other models focused on long-term behaviors.
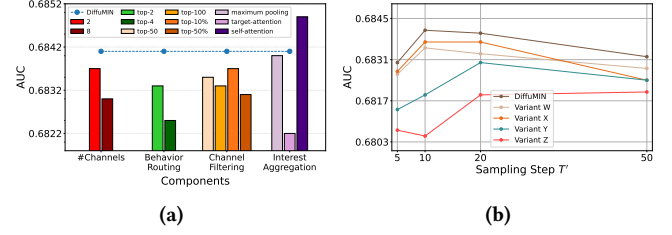
4.4 Analysis of OMIE (RQ3)

In this section, we conduct experiments on the industrial dataset to analyze the OMIE module by constructing various variants. We adjust the number of channels, behavior routing, channel filtering, and interest aggregation, using DiffuMIN configured with {4, top-1, top-20%, mean pooling} as the baseline.

Figure 2a shows that selecting an optimal number of channels is critical; too few channels limit interest diversity, while too many can reduce the effectiveness of interest channel capabilities. For behavior routing, the top-1 routing method minimizes redundancy within channels, thereby boosting performance. When filtering channels, the top-$p\%$ strategy is superior to top-k because of the varying behavior counts across channels. Regarding interest aggregation, further aggregation using target-attention does not yield additional benefits since the relationship between the target and behaviors is already established. Although self-attention can enhance performance, its significant time demands pose challenges for online deployment.

4.5 Analysis of DMIG (RQ4)

4.5.1 Advantages of Diffusion Module. Currently, GANs [70] and VAEs [71] are prevalent generative modeling methods. Compared to GANs, VAEs are more extensively applied in recommender systems [68, 69]. Thus, we focus on examining the advantages of diffusion models over VAEs in this section.

Theoretically, VAEs utilize a variational posterior to approximate the true posterior. The model's effectiveness diminishes if the variational posterior is overly simplistic, while optimization becomes challenging if the posterior is too complex. Furthermore, VAEs simultaneously optimize both the conditional distribution and the variational posterior, resulting in a large search space and issues such as posterior collapse. In contrast, diffusion models first define the variational posterior through a Markov Chain and subsequently fit it using the conditional distribution, thereby focusing solely on optimizing the conditional distribution. As a result, diffusion models provide superior and more stable performance.

(a)   **(b)**

Figure 2: Comparative analysis on the industrial dataset: (a) variants of OMIE module, (b) variants of DMIG module.

Table 4: Performances with different generative modeling methods on three datasets.

Models	Industry	Alibaba	Ele.me
DiffuMIN	0.6841	0.6282	0.6462
DiffuMIN-VAE	0.6830	0.6277	0.6444

Experimentally, we refer to CVAE[3] and ContrastVAE[4][68] to implement the VAE module, replacing the diffusion module in DiffuMIN to create the variant DiffuMIN-VAE. The experimental results, presented in Table 4, indicate that DiffuMIN-VAE does not achieve superior results on the industry and Ele.me datasets, and only approximates the results of DiffuMIN on the Alibaba dataset.

4.5.2 Adaptation of Diffusion Module. In our model, multiple enhancements have been applied to the traditional diffusion model. Specifically, we propose guiding the diffusion module using contextual interests and interest channels and utilizing perturbed user interests as the starting point for generation. To evaluate the effectiveness of our approach, we construct variants W, X, Y, and Z: Variant W lacks contextual interests guidance, variant X lacks interest channels guidance, variant Y omits them both, and variant Z generates directly from Gaussian noise.

The results, shown in Figure 2b, indicate that while the interests derived from OMIE offer personalized information for both user and target, removing contextual interests or interest channels guidance harms the generation quality, especially when both are absent. Variant Z's results highlight the difficulties in generating high-quality embeddings directly from Gaussian noise, necessitating additional sampling steps.

4.6 Case Study (RQ5)

To visually assess how DiffuMIN better preserves and explores user interest spaces compared to traditional models for long-term behaviors, we apply t-SNE to visualize the embeddings of various user interests. Figure 3 depicts two cases from the industrial dataset, each containing a user and the target clicked by the user. Large pentagrams denote targets, while small pentagrams signify interest channels decomposed from the target by using DiffuMIN. Large and small colored circles are the aggregated and augmented interests in DiffuMIN, respectively. Gray circles denote user interests derived from TWIN's first stage.

---

[3] https://github.com/mingukkang/CVAE
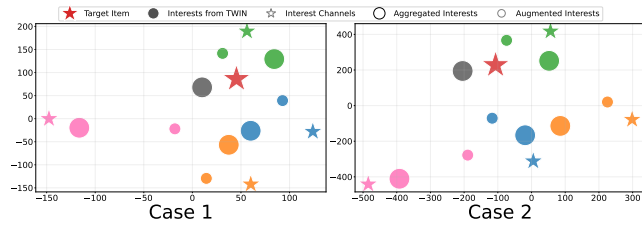[4] https://github.com/YuWang-1024/ContrastVAE

Figure 3: Case study on two specific users.

As illustrated in Figure 3, each gray circle is close to a large pentagram, indicating the TWIN captures only user interest that is close to the target. However, each small pentagram is close to one or two colored circles, indicating that, in DiffuMIN, the target matches the user's interests. The reason is as follows. Our model employs the orthogonal decomposition of the target to achieve finer-grained representations as interest channels. This strategy allows us to select and aggregate user behaviors according to distinct interest channels, effectively preserving a diverse range of user interests. Furthermore, our approach extends these interests through generative methods, significantly broadening the exploration of the user's interest space.

4.7 Online A/B Test (RQ6)

We conducted an online A/B test deploying our DiffuMIN model in the live production environment for 7 days. The baseline model, serving the list advertisement scenario, integrates foundational models like DIN and SIM and has undergone several iterations. In the A/B test, we replaced SIM in the baseline with DiffuMIN. This led to a 1.52% increase in CTR and a 1.10% increase in CPM, with inference time rising slightly from 33ms to 35ms.

5 CONCLUSION

In this paper, we propose the DiffuMIN model for effectively modeling long-term user behaviors. We begin by proposing a target-oriented multi-interest extraction method to disentangle and extract multiple user interests. This is complemented by a diffusion module, guided by contextual interests and interest channels, to generate multiple augmented interests. Our approach significantly preserves and expands the constrained interest space in long-term behavior modeling, thereby enhancing the overall capacity of the model. Results from offline experiments and online A/B testing demonstrate the superiority of DiffuMIN over existing models.

REFERENCES

[1] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep interest network for click-through rate prediction," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1059–1068.
[2] Y. Cao, X. Zhou, J. Feng, P. Huang, Y. Xiao, D. Chen, and S. Chen, "Sampling is all you need on modeling long-term user behaviors for ctr prediction," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2974–2983.
[3] Q. Pi, G. Zhou, Y. Zhang, Z. Wang, L. Ren, Y. Fan, X. Zhu, and K. Gai, "Search-based user interest modeling with lifelong sequential behavior data for click-through rate prediction," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2685–2692.
[4] Q. Chen, C. Pei, S. Lv, C. Li, J. Ge, and W. Ou, "End-to-end user behavior retrieval in click-through rateprediction model," *arXiv preprint arXiv:2108.04468*, 2021.
[5] W. Bian, K. Wu, L. Ren, Q. Pi, Y. Zhang, C. Xiao, X.-R. Sheng, Y.-N. Zhu, Z. Chan, N. Mou *et al.*, "Can: Feature co-action for click-through rate prediction," *arXiv preprint arXiv:2011.05625*, 2020.
[6] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir *et al.*, "Wide & deep learning for recommender systems," in *Proceedings of the 1st workshop on deep learning for recommender systems*, 2016, pp. 7–10.
[7] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "Deepfm: a factorization-machine based neural network for ctr prediction," *arXiv preprint arXiv:1703.04247*, 2017.
[8] J. Xiao, H. Ye, X. He, H. Zhang, F. Wu, and T.-S. Chua, "Attentional factorization machines: Learning the weight of feature interactions via attention networks," *arXiv preprint arXiv:1708.04617*, 2017.
[9] Y. Yang, B. Xu, S. Shen, F. Shen, and J. Zhao, "Operation-aware neural networks for user response prediction," *Neural Networks*, vol. 121, pp. 161–168, 2020.
[10] Y. Zhang, T. Shi, F. Feng, W. Wang, D. Wang, X. He, and Y. Zhang, "Reformulating ctr prediction: Learning invariant feature interactions for recommendation," *arXiv preprint arXiv:2304.13643*, 2023.
[11] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, and K. Gai, "Deep interest evolution network for click-through rate prediction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 5941–5948.
[12] Y. Feng, F. Lv, W. Shen, M. Wang, F. Sun, Y. Zhu, and K. Yang, "Deep session interest network for click-through rate prediction," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 2301–2307.
[13] Q. Chen, H. Zhao, W. Li, P. Huang, and W. Ou, "Behavior sequence transformer for e-commerce recommendation in alibaba," in *Proceedings of the 1st international workshop on deep learning practice for high-dimensional sparse data*, 2019, pp. 1–4.
[14] Q. Pi, W. Bian, G. Zhou, X. Zhu, and K. Gai, "Practice on long sequential user behavior modeling for click-through rate prediction," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2671–2679.
[15] J. Qin, W. Zhang, X. Wu, J. Jin, Y. Fang, and Y. Yu, "User behavior retrieval for click-through rate prediction," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2347–2356.
[16] K. Ren, J. Qin, Y. Fang, W. Zhang, L. Zheng, W. Bian, G. Zhou, J. Xu, Y. Yu, X. Zhu *et al.*, "Lifelong sequential modeling with personalized memorization for user response prediction," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 565–574.
[17] G. S. Manku, A. Jain, and A. Das Sarma, "Detecting near-duplicates for web crawling," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 141–150.
[18] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proceedings of the 10th ACM conference on recommender systems*, 2016, pp. 191–198.
[19] J. Lian, X. Zhou, F. Zhang, Z. Chen, X. Xie, and G. Sun, "xdeepfm: Combining explicit and implicit feature interactions for recommender systems," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1754–1763.
[20] Y. Qu, H. Cai, K. Ren, W. Zhang, Y. Yu, Y. Wen, and J. Wang, "Product-based neural networks for user response prediction," in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 1149–1154.
[21] R. Wang, B. Fu, G. Fu, and M. Wang, "Deep & cross network for ad click predictions," in *Proceedings of the ADKDD'17*, 2017, pp. 1–7.
[22] C. Li, Z. Liu, M. Wu, Y. Xu, H. Zhao, P. Huang, G. Kang, Q. Chen, W. Li, and D. L. Lee, "Multi-interest network with dynamic routing for recommendation at tmall," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 2615–2623.
[23] L. Yan, W.-J. Li, G.-R. Xue, and D. Han, "Coupled group lasso for web-scale ctr prediction in display advertising," in *International conference on machine learning*. PMLR, 2014, pp. 802–810.
[24] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 197–206.
[25] Z. Si, L. Guan, Z. Sun, X. Zang, J. Lu, Y. Hui, X. Cao, Z. Yang, Y. Zheng, D. Leng *et al.*, "Twin v2: Scaling ultra-long user behavior sequence modeling for enhanced ctr prediction at kuaishou," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 4890–4897.
[26] C. Liu, X. Li, G. Cai, Z. Dong, H. Zhu, and L. Shang, "Noninvasive self-attention for side information fusion in sequential recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4249–4256.
[27] Y. Xie, P. Zhou, and S. Kim, "Decoupled side information fusion for sequential recommendation," in *Proceedings of the 45th International ACM SIGIR Conference

on Research and Development in Information Retrieval, 2022, pp. 1611–1621.

[28] T. Zhang, P. Zhao, Y. Liu, V. S. Sheng, J. Xu, D. Wang, G. Liu, X. Zhou *et al.*, "Feature-level deeper self-attention network for sequential recommendation." in *IJCAI*, 2019, pp. 4320–4326.

[29] K. Zhou, H. Wang, W. X. Zhao, Y. Zhu, S. Wang, F. Zhang, Z. Wang, and J.-R. Wen, "S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 1893–1902.

[30] W. Guo, C. Zhang, Z. He, J. Qin, H. Guo, B. Chen, R. Tang, X. He, and R. Zhang, "Miss: Multi-interest self-supervised learning framework for click-through rate prediction," in *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 2022, pp. 727–740.

[31] J. Chang, C. Zhang, Z. Fu, X. Zang, L. Guan, J. Lu, Y. Hui, D. Leng, Y. Niu, Y. Song *et al.*, "Twin: Two-stage interest network for lifelong user behavior modeling in ctr prediction at kuaishou," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 3785–3794.

[32] A. Rashed, S. Elsayed, and L. Schmidt-Thieme, "Context and attribute-aware sequential recommendation via cross-attention," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 71–80.

[33] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," *arXiv preprint arXiv:2303.09472*, 2023.

[34] S. Rendle, "Factorization machines," in *2010 IEEE International conference on data mining*. IEEE, 2010, pp. 995–1000.

[35] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers *et al.*, "Practical lessons from predicting clicks on ads at facebook," in *Proceedings of the eighth international workshop on data mining for online advertising*, 2014, pp. 1–9.

[36] H. Chung, B. Sim, and J. C. Ye, "Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 413–12 422.

[37] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.

[38] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[39] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[40] X. Lin, X. Chen, C. Wang, H. Shu, L. Song, B. Li, and P. Jiang, "Discrete conditional diffusion for reranking in recommendation," in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 161–169.

[41] W. Xie, H. Wang, L. Zhang, R. Zhou, D. Lian, and E. Chen, "Breaking determinism: Fuzzy modeling of sequential recommendation using discrete state space diffusion model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 22 720–22 744, 2024.

[42] Q. Li, H. Ma, W. Jin, Y. Ji, and Z. Li, "Multi-interest network with simple diffusion for multi-behavior sequential recommendation," in *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 2024, pp. 734–742.

[43] W. Xie, R. Zhou, H. Wang, T. Shen, and E. Chen, "Bridging user dynamics: Transforming sequential recommendations with schrödinger bridge and diffusion models," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 2618–2628.

[44] J. Lin, J. Liu, J. Zhu, Y. Xi, C. Liu, Y. Zhang, Y. Yu, and W. Zhang, "A survey on diffusion models for recommender systems," *arXiv preprint arXiv:2409.05033*, 2024.

[45] T.-R. Wei and Y. Fang, "Diffusion models in recommendation systems: A survey," *arXiv preprint arXiv:2501.10548*, 2025.

[46] W. Zhu, L. Wang, and J. Wu, "Addressing cold-start problem in click-through rate prediction via supervised diffusion modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 13 455–13 463.

[47] Z. Feng, J. Xie, K. Li, Y. Qin, P. Wang, Q. Li, B. Yin, X. Li, W. Lin, and S. Wang, "Context-based fast recommendation strategy for long user behavior sequence in meituan waimai," in *Companion Proceedings of the ACM Web Conference 2024*, 2024, pp. 355–363.

[48] Z. Liu, J. Lian, J. Yang, D. Lian, and X. Xie, "Octopus: Comprehensive and elastic user representation for the generation of recommendation candidates," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 2020, pp. 289–298.

[49] W. Guo, C. Zhang, Z. He, J. Qin, H. Guo, B. Chen, R. Tang, X. He, and R. Zhang, "Miss: Multi-interest self-supervised learning framework for click-through rate prediction," in *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 2022, pp. 727–740.

[50] H. Jiang, W. Wang, Y. Wei, Z. Gao, Y. Wang, and L. Nie, "What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation," in *Proceedings of the 28th ACM International conference on Multimedia*, 2020, pp. 3487–3495.

[51] Y. Xie, J. Gao, P. Zhou, Q. Ye, Y. Hua, J. B. Kim, F. Wu, and S. Kim, "Rethinking multi-interest learning for candidate matching in recommender systems," in *Proceedings of the 17th ACM conference on recommender systems*, 2023, pp. 283–293.

[52] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[53] J. Zhai, L. Liao, X. Liu, Y. Wang, R. Li, X. Cao, L. Gao, Z. Gong, F. Gu, J. He *et al.*, "Actions speak louder than words: trillion-parameter sequential transducers for generative recommendations," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 58 484–58 509.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[55] B. Li, B. Jin, Y. Yu, Y. Zheng, J. Song, W. Zhuo, and T. Xiang, "Orthogonal hyper-category guided multi-interest elicitation for micro-video matching," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[56] B. Li, B. Jin, J. Song, Y. Yu, Y. Zheng, and W. Zhou, "Improving micro-video recommendation via contrastive multiple interests," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 2377–2381.

[57] Z. Xiao, L. Yang, W. Jiang, Y. Wei, Y. Hu, and H. Wang, "Deep multi-interest network for click-through rate prediction," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 2265–2268.

[58] J. Yan, L. Jiang, J. Cui, Z. Zhao, X. Bin, F. Zhang, and Z. Liu, "Trinity: Syncretizing multi-/long-tail/long-term interests all in one," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6095–6104.

[59] Y. Cen, J. Zhang, X. Zou, C. Zhou, H. Yang, and J. Tang, "Controllable multi-interest framework for recommendation," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 2942–2951.

[60] L. Liu, L. Cai, C. Zhang, X. Zhao, J. Gao, W. Wang, Y. Lv, W. Fan, Y. Wang, M. He *et al.*, "Linrec: Linear attention mechanism for long-term sequential recommender systems," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 289–299.

[61] H. Ma, R. Xie, L. Meng, Y. Yang, X. Sun, and Z. Kang, "Seedrec: sememe-based diffusion for sequential recommendation," in *Proceedings of IJCAI*, 2024, pp. 1–9.

[62] Z. Cui, H. Wu, B. He, J. Cheng, and C. Ma, "Context matters: Enhancing sequential recommendation with context-aware diffusion-based contrastive learning," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 404–414.

[63] Z. Wu, X. Wang, H. Chen, K. Li, Y. Han, L. Sun, and W. Zhu, "Diff4rec: Sequential recommendation with curriculum-scheduled diffusion augmentation," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 9329–9335.

[64] H. Ma, R. Xie, L. Meng, X. Chen, X. Zhang, L. Lin, and Z. Kang, "Plug-in diffusion model for sequential recommendation," *arXiv preprint arXiv:2401.02913*, 2024.

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

[66] W. Wang, Y. Xu, F. Feng, X. Lin, X. He, and T.-S. Chua, "Diffusion recommender model," *arXiv preprint arXiv:2304.04971*, 2023.

[67] Z. Yang, J. Wu, Z. Wang, X. Wang, Y. Yuan, and X. He, "Generate what you prefer: Reshaping sequential recommendation via guided diffusion," *arXiv preprint arXiv:2310.20453*, 2023.

[68] Y. Wang, H. Zhang, Z. Liu, L. Yang, and P. S. Yu, "Contrastvae: Contrastive variational autoencoder for sequential recommendation," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2056–2066.

[69] J. Zhao, P. Zhao, L. Zhao, Y. Liu, V. S. Sheng, and X. Zhou, "Variational self-attention network for sequential recommendation," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 1559–1570.

[70] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[71] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.