

# HoMer: Addressing Heterogeneities by Modeling Sequential and Set-wise Contexts for CTR Prediction

Shuwei Chen\*, Jiajun Cui\*, Zhengqi Xu\*, Fan Zhang, Jiangke Fan<sup>†</sup>, Teng Zhang, Xingxing Wang  
 {chenshuwei04,cuijiajun02,xuzhengqi02,zhangfan133,jiangke.fan,zhangteng09,wangxingxing04}@meituan.com  
 Meituan  
 Shanghai, China

## Abstract

Click-through rate (CTR) prediction, which models behavior sequence and non-sequential features (e.g., user/item profiles or cross features) to infer user interest, underpins industrial recommender systems. However, most methods face three forms of **heterogeneity** that degrade predictive performance: (i) **Feature Heterogeneity** persists when limited sequence side features provide less granular interest representation compared to extensive non-sequential features, thereby impairing sequence modeling performance; (ii) **Context Heterogeneity** arises because a user's interest in an item will be influenced by other items, yet point-wise prediction neglects cross-item interaction context from the entire item set; (iii) **Architecture Heterogeneity** stems from the fragmented integration of specialized network modules, which compounds the model's effectiveness, efficiency and scalability in industrial deployments. To tackle the above limitations, we propose **HoMer**, a Homogeneous-Oriented Transformer for modeling sequential and set-wise contexts. First, we align sequence side features with non-sequential features for accurate sequence modeling and fine-grained interest representation. Second, we shift the prediction paradigm from point-wise to set-wise, facilitating cross-item interaction in a highly parallel manner. Third, HoMer's unified encoder-decoder architecture achieves dual optimization through structural simplification and shared computation, ensuring computational efficiency while maintaining scalability with model size. Without arduous modification to the prediction pipeline, HoMer successfully scales up and outperforms our industrial baseline by 0.0099 in the AUC metric, and enhances online business metrics like CTR/RPM by 1.99%/2.46%. Additionally, HoMer saves 27% of GPU resources via preliminary engineering optimization, further validating its superiority and practicality.

## CCS Concepts

- Information systems → Recommender systems; Online advertising.

\* Equal contribution.

<sup>†</sup>Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/2018/06

<https://doi.org/XXXXXXX.XXXXXXX>

## Keywords

CTR Prediction, Sequence Modeling, Item Set Context, Transformer

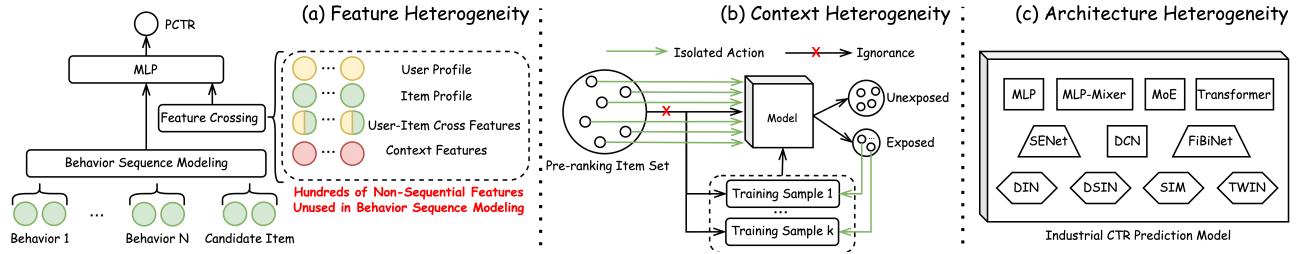
### ACM Reference Format:

Shuwei Chen\*, Jiajun Cui\*, Zhengqi Xu\*, Fan Zhang, Jiangke Fan<sup>†</sup>, Teng Zhang, Xingxing Wang. 2018. HoMer: Addressing Heterogeneities by Modeling Sequential and Set-wise Contexts for CTR Prediction. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 Introduction

Click-through rate (CTR) prediction refers to the task of estimating the probability of a user clicking on a given item. It has a direct impact on user engagement and platform revenue, and plays a foundational role in industrial recommender systems. With the rise of deep learning [14], Deep Learning Recommendation Models (DLRMs) have been widely proposed for CTR prediction [3, 4, 7, 8, 19]. Among these developments, behavior sequence modeling [1, 21, 24, 43, 45] and feature crossing [7, 15, 25, 28, 29] are two of the most popular topics. Specifically, the former matches items with a user's behavior sequence to extract dynamic user interest representation, where several side features (e.g., item category and price) are incorporated to enhance sequence representation capability [16, 32, 38]. The latter characterizes user interest representation by crossing extensive non-sequential features, including user/item profiles, user-item cross features and context features. Recent years have witnessed significant advancements in these methods, substantially enhancing CTR prediction performance.

However, we contend that three persistent forms of heterogeneity adversely affect CTR prediction performance: (i) **Feature Heterogeneity**: As illustrated in Fig. 1(a), both behavior sequence and non-sequential features are utilized to infer CTR. The extensive non-sequential features are capable for representing fine-grained user interest toward an item, but each behavior is only represented by several side features, producing coarse-grained interest representation. (ii) **Context Heterogeneity**: As demonstrated in Fig. 1(b), the upstream pre-ranking stage generates a set of items (typically numbering in tens to hundreds). The CTR prediction model subsequently predicts CTR for these items in a point-wise paradigm, after which the top-ranked items will be exposed to users on the same display page. However, a user's interest in any item may be influenced by other co-exposed items within the same page view [2, 40]. For instance, when multiple similar items are exposed concurrently, users tend to engage in comparative evaluation before making click decisions. The isolated prediction paradigm fundamentally disregards the cross-item interaction context, thereby creating a misalignment with authentic user behavior patterns. (iii) **Architecture**



**Figure 1: Illustrations of heterogeneities in traditional CTR prediction paradigm.** (a) **Feature Heterogeneity:** The misalignment between sequence side features and non-sequential features produces coarse-grained user interest representation. (b) **Context Heterogeneity:** In point-wise prediction paradigm, the neglect of cross-item interaction context from the entire item set limits the model’s capability to capture authentic user behavior patterns. (c) **Architecture Heterogeneity:** The fragmented integration of specialized network modules constrains the model’s effectiveness, efficiency and scalability.

**Heterogeneity:** As shown in Fig. 1(c), industrial CTR prediction models undergo continuous iteration and integration across multiple technical dimensions, resulting in architectures that inherently contain heterogeneous modules. For example, DIN [45], DSIN [6] and TWIN [1] for behavior sequence modeling; SENet [10] and FiBiNet [11] for attention-based gating; and DCN [28] and MoE [23] for feature enhancement. These modules exhibit overlap or a seesaw effectiveness, with some even becoming ineffective. Besides, engineering optimization for a single module is unlikely to significantly enhance overall model efficiency. All these factors complicate model maintenance and iteration, not to mention its potential for scaling up [27, 35, 41, 42, 46] in the era of computing.

We propose **Homogeneous-Oriented Transformer (HoMer)**, which models sequential and set-wise contexts in a computationally efficient manner, to tackle the above heterogeneities. First, HoMer aligns sequence side features with non-sequential features for fine-grained user interest representation. It means that each behavior now contains complete user/item profiles, user-item cross features and context features recorded on the corresponding historical request. For convenience, this paper refers to the constructed sequence as **panoramic sequence**, as it incorporates all features within the user lifecycle that are perceivable during the CTR prediction stage. Second, traditional CTR prediction methods construct one sample for each item, and invoke model prediction for each sample in isolation. HoMer shifts the prediction paradigm from point-wise to set-wise by aggregating the non-sequential features of all the items into one sample, which enables it to perform cross-item interaction and parallel prediction for all items in a single model invocation. Third, HoMer adopts a unified homogeneous-oriented encoder-decoder architecture [26], where the encoder is responsible for extracting fine-grained user interest representation from the panoramic sequence, and the decoder is responsible for cross-item and user-item interactions. Under the above settings, HoMer ingeniously tackles the three forms of heterogeneity, and the computational overhead introduced by the panoramic sequence is naturally shared by all the items, keeping HoMer’s efficiency.

Moreover, HoMer also shows great practicality from multiple dimensions. Traditional methods generally construct one offline sample for each exposed item, resulting in redundant storage for behavior sequence. Instead, HoMer constructs one offline sample

for each request, which not only reduces storage cost, but also saves data I/O during model training. To prevent overfitting and reduce training cost, traditional methods usually performs negative sampling on the isolated exposed corpus. However, this approach inherently hinders the model from understanding rich and authentic user behavior patterns. By contrast, HoMer is capable of efficiently consuming the entire training corpus to model more precise patterns, thereby improving prediction performance. HoMer also simplifies online prediction service. For example, HoMer’s single-pass processing of panoramic sequence inherently eliminates the need for deduplication mechanisms in embedding lookup operations and associated computations.

We conduct extensive offline and online experiments in the search advertising scenario of Meituan<sup>1</sup>. The experimental results demonstrate that HoMer outperforms our industrial baseline by 0.0099 in the Area Under Curve (AUC) metric. It should be noted that a 0.001 AUC improvement is considered significant and worth deploying in industrial recommender systems. Deployed across Meituan’s online traffic from tens of millions of users, HoMer achieves a 1.99% lift in CTR and a 2.46% increase in Revenue Per Mille (RPM). Notably, even with only preliminary kernel fusion optimizations, HoMer has elevated online Model FLOPs Utilization (MFU) from 7.8% to 12.2%, while simultaneously reducing online GPU resource consumption by 27%. Furthermore, the streamlined architecture and computational efficiency of HoMer enable seamless scalability, yielding significant performance improvements (as discussed in Section 5).

The key contributions of this paper are summarized as follows:

- **Motivation:** We identify three heterogeneity forms (feature, context, architecture) that degrade CTR prediction performance.
- **Methodology:** We propose HoMer, an unified, efficient and practical transformer, to tackle the above heterogeneities. It jointly models panoramic sequence for fine-grained user interest representation, and set-wise cross-item interaction context for authentic user behavior patterns learning.

<sup>1</sup>One of China’s largest platforms providing local lifestyle services.

- **Experiment:** We conduct extensive experiments in the search advertising scenario of Meituan. Both offline and online experimental results not only consistently validate HoMer’s superiority and efficiency, but also clearly demonstrate its scaling potential.

## 2 Related Work

### 2.1 Deep Learning Recommendation Models

Deep learning has become the cornerstone of industrial CTR prediction, with models continually advancing in their capacity to capture feature interactions and user behavior pattern. The Deep Learning Recommendation Model (DLRM) [20] pioneered the integration of embedding layers and multilayer perceptrons (MLPs) to represent complex relationships among features. Building on this, DeepFM [7] and Wide & Deep [3] combined factorization machines and linear models with deep neural networks to capture both low- and high-order interactions, balancing memorization and generalization. xDeepFM [15] further improved modeling capabilities via explicit and implicit interaction mechanisms. Afterwards, user behavior modeling remains central to CTR prediction. DIN [45] introduced attention-based mechanisms to extract user interests from historical actions, while DIEN [44] leveraged recurrent networks to capture the evolving nature of user interests. Self-attention techniques, as in AutoInt [25], have also proven effective for modeling behavior patterns and context. Recent works like Wukong [36] highlight the scalability of large unified models, and CIM [40] enhances user modeling by incorporating candidate set information. Despite their successes, these models mostly adhere to a point-wise paradigm and have grown increasingly complex through iterative development, resulting in industrial systems that are challenging to optimize and scale.

### 2.2 Transformer in Recommendation

Transformer architectures have recently gained prominence in recommendation systems due to their strength in modeling complex user-item interactions and long-range dependencies. Their flexibility supports both discriminative and generative recommendation tasks. Recent research has explored generative paradigms powered by Transformers, aiming to accelerate inference and optimize auto-regressive generation in large-scale scenarios, as seen in EARN [33] and Act-With-Think [30]. Unified frameworks like OneRec [41] integrate retrieval and ranking for end-to-end recommendation, while generative pretraining techniques [27] have improved industrial performance. Efforts to enhance generative recommenders include advanced tokenization methods, such as ActionPiece [9] and learnable item tokenization [17], which support context-aware and end-to-end generative modeling. Research on unifying retrieval and ranking [37] and supporting multi-behavior recommendation [18] further expand the scope of Transformer-based models. EAGER [31] demonstrates holistic list optimization, while collaborative semantics [39] and sparse-dense representations [34] showcase the paradigm’s scalability. Notably, HSTU [35] pushes the boundaries with trillion-parameter sequential transducers, and the generative retrieval paradigm [22] is reshaping the integration of retrieval and ranking. Nowadays, transformer-based models support joint inference over multiple candidates are mainstream. Our proposed HoMer does not only continue this trend, but

also offering a unified, scalable architecture built from the ground up for end-to-end recommendation.

## 3 Preliminary

A CTR prediction model  $\mathcal{M}$  estimates the probability of a user  $u \in \mathcal{U}$  clicking on an item  $i \in \mathcal{I}$  using the following features:

- **User Profile**  $f_u$  denotes a user’s basic attributes, e.g., age, gender and career.
- **Item Profile**  $f_i$  denotes an item’s static properties (e.g., category, price and rating) and dynamic metrics (e.g., CTR and popularity).
- **User-Item Cross Features**  $f_{u,i}$  captures crossing signals between user  $u$  and item  $i$ , such as historical click/purchase frequency and affinity scores.
- **Context Features**  $f_c$  stands for real-time environmental factors influencing user decisions, such as time of day, location and device type.
- **Behavior Sequence**  $f_{seq} = \{f_{seq,1}, f_{seq,2}, \dots, f_{seq,n}\}$  records a user’s  $n$  historical interactions with items (e.g., impressions, clicks, and orders), and will be leveraged to model user interest representation. In most methods, several side features (e.g., item category and price) are incorporated to each behavior to enhance sequence modeling, i.e.,  $f_{seq,i} = \{f_{seq,i}^1, f_{seq,i}^2, \dots, f_{seq,i}^s\}$ , where  $s$  denotes the number of side features. Since the length of behavior sequence is typically long, to maintain data storage and model efficiency, the inequality  $s \ll |f_u| + |f_i| + |f_{u,i}| + |f_c|$  generally holds in point-wise prediction paradigm. It indicates that sequence side features provide less granular interest representation compared to the non-sequential features.

Following the above notations, the point-wise CTR prediction can be formulated as:

$$p_{u,i} = \mathcal{M}(f_u, f_i, f_{u,i}, f_c, f_{seq}), \quad (1)$$

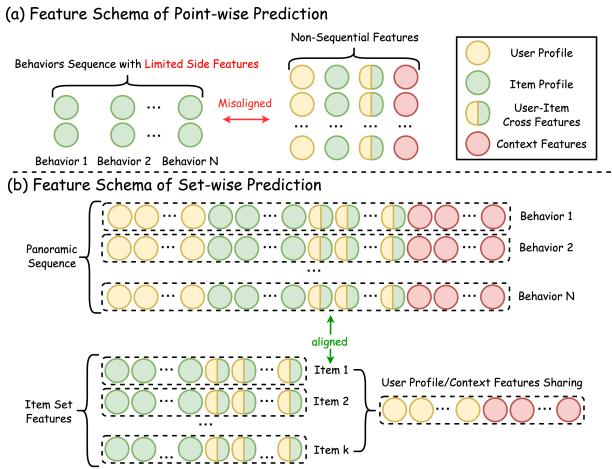
where  $p_{u,i}$  is the predicted CTR of user  $u$  on item  $i$ . This paper proposes a set-wise prediction method, HoMer, which can be formulated as:

$$p_{u,i_1}, \dots, p_{u,i_k} = \text{HoMer}(f_u, f_c, f_{pan\_seq}, f_{i_1}, f_{u,i_1}, \dots, f_{i_k}, f_{u,i_k}), \quad (2)$$

where  $k$  is the number of items in the set, and  $p_{u,i_k}$  represents the predicted CTR of the  $k$ -th item. The term  $f_{pan\_seq}$  refers to our proposed panoramic sequence, designing to learn fine-grained user interest representation. The terms  $f_{i_k}$  and  $f_{u,i_k}$  correspond to the item profile and user-item cross features of the  $k$ -th item, which are responsible for modeling cross-item and user-item interactions. As shown in the equation, HoMer performs set-wise prediction within a single model invocation, ensuring computational efficiency.

## 4 Methodology

Our analysis identifies three inherent heterogeneities in traditional point-wise CTR prediction methods. This section presents how HoMer addresses these heterogeneities and improves CTR prediction performance. First, to tackle feature heterogeneity, HoMer transforms the behavior sequence into panoramic sequence by aligning limited side features with non-sequential features, thereby enabling fine-grained user interest representation. Second, to address context heterogeneity, HoMer shifts the prediction paradigm from point-wise to set-wise, which facilitates modeling of



**Figure 2: Comparison of feature schemas between point-wise prediction and HoMer’s set-wise prediction.**

cross-item interactions and authentic user behavior patterns. Third, to mitigate architecture heterogeneity, HoMer adopts a unified homogeneous-oriented transformer architecture that incorporates the aforementioned optimizations while maintaining efficiency, practicality, and scalability. The section concludes with implementation details on model training and deployment.

#### 4.1 Panoramic Sequence

Fig. 2(a) illustrates the feature schema of point-wise prediction, which consists of a user’s behavior sequence and non-sequential features. The former contains  $N$  behaviors corresponding to  $N$  items, each of which is represented by several side features, such as item category and price. The latter contains extensive features that characterize the user’s fine-grained interest representation toward the item. To capture user’s dynamic interest representation toward the item, most CTR prediction methods play matching between the item and the behavior sequence. However, due to the limited information provided by the several side features, behavior sequence produces coarse-grained user interest representation, resulting in insufficient improvement in CTR prediction performance.

We enrich the behavior sequence by aligning side features with non-sequential features to mitigate feature heterogeneity. The structure of the resulted panoramic sequence is illustrated in the upper panel of Fig. 2(b). For each behavior in panoramic sequence, its side features consist of all the non-sequential features recorded in the corresponding historical request, which can precisely describe the user’s decision context at that moment. In this way, the panoramic sequence contains all the features within the user lifecycle, benefitting behavior sequence modeling and fine-grained user interest representation. It is important to note that the panoramic sequence can be quickly constructed by chronologically aggregating a user’s historical CTR prediction samples. Moreover, the online log service can be adapted for real-time panoramic sequence construction.

#### 4.2 Set-wise CTR Prediction

The point-wise CTR prediction paradigm operates through two distinct phases: (1) Offline training phase converts impression logs into discrete training samples using strategically sampled negatives; and (2) Online serving phase executes independent prediction for individual items. This presents the following limitations:

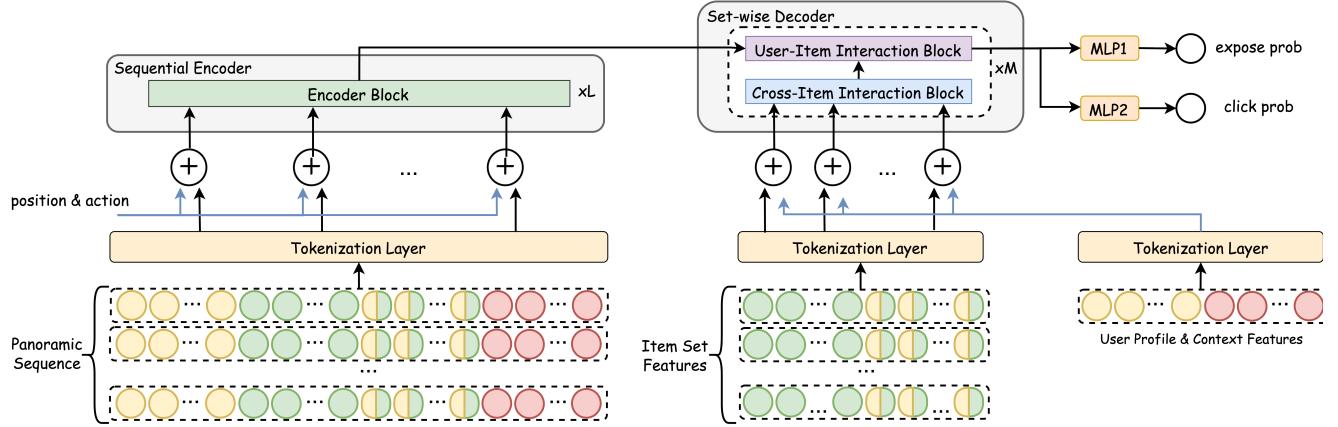
- **Storage Inefficiency:** Multiple training samples generated from single page views contain redundant features (identical user profile, context features, and behavior sequence), resulting in inefficient storage utilization, particularly when implementing our proposed panoramic sequence.
- **Sampling Bias:** Negative sampling strategies exclusively utilize exposed items, while systematically excluded unexposed items contain valuable implicit signals for user interest modeling.
- **Contextual Blindness:** The single-item focus of training samples ignores cross-item interaction context within co-exposed page, leading to compromised understanding of authentic user behavior and ultimately constrains model performance.
- **Computational Redundancy:** The serving phase incurs duplicated computations (e.g., table lookup operations and network computations for user profile, context features and behavior sequence) across items from the same request, necessitating deduplication mechanisms to maintain online inference efficiency.

In this paper, we introduce a set-wise CTR prediction paradigm designed to address the aforementioned limitations simultaneously. This paradigm is mathematically formulated in Eq. (2), with the corresponding feature schema depicted in the lower panel of Fig. 2(b). Instead of generating a sample for each item, we construct a single sample for every request. For items within the same request, their profiles and user-item cross features are aggregated collectively. Item-independent features, such as user profiles, context features, and panoramic sequence, are shared across all items within the request, thereby reducing storage inefficiency. In HoMer, items interact with each other to capture authentic user behavior patterns, thus addressing the issues of sampling bias and contextual blindness. By computing item-independent features only once and sharing them across all items, HoMer facilitates efficient cross-item interaction and parallel prediction for all items through a single model invocation. Noting that the feature schemas of panoramic sequence and item set are aligned.

#### 4.3 Homogenous-Oriented Transformer

Point-wise CTR prediction models typically consist of heterogeneous modules and perform prediction independently for each item. As discussed in previous sections, this paradigm exhibits inefficiency and suboptimal performance. In this paper, we introduce HoMer, a unified Homogenous-Oriented Transformer designed to capture both sequential and set-wise contexts while maintaining efficiency, practicality, and scalability. Fig. 4 gives an overview of HoMer, which includes a sequential encoder responsible for modeling fine-grained user interest representation, and a set-wise decoder tasked with capturing cross-item interaction context.

**4.3.1 Sequential Encoder.** The sequential encoder is composed of a series of encoder blocks, and processes panoramic sequence to generate fine-grained user interest representation. Each behavior,



**Figure 3: The overall architecture of HoMer.** The sequential encoder is responsible for modeling fine-grained user interest representation from panoramic sequence, and the set-wise decoder is tasked with capturing cross-item interaction context from the features of the entire item set.

along with its extensive side features, is passed through a shared tokenization layer and projected into a  $d$ -dimensional behavior embedding. After combining these embeddings with the positional and actional embeddings of the behaviors, we obtain the initial user interest representation. This procedure can be formulated as:

$$E_i^0 = \sigma_1(\phi(f_{pan\_seq,i})) + \sigma_2(\phi(p_i)) + \sigma_3(\phi(a_i)) \quad (3)$$

where  $E_i^0 \in \mathbb{R}^d$  is the initial user interest represented by the  $i$ -th behavior. Here,  $f_{pan\_seq,i}$  refers to the  $i$ -th behavior and its extensive side features in the panoramic sequence, while  $p_i$  and  $a_i$  denote the position and user action (e.g., click or not) of the  $i$ -th behavior. The term  $\phi$  denotes embedding lookup operation, and  $\sigma_1$ ,  $\sigma_2$  and  $\sigma_3$  are the tokenization layers, which are implemented by a fully connected layer.

Subsequently, the initial representation  $E^0 = \{E_1^0, E_2^0, \dots, E_N^0\}$  is fed into a stack of encoder block to generate fine-grained user interest representation, which can be formulated as:

$$Q^l = \tau(\sigma_Q^l(E^{l-1})), K^l = \tau(\sigma_K^l(E^{l-1})), V^l = \tau(\sigma_V^l(E^{l-1})) \quad (4)$$

$$E^l = \omega^l(\sigma^l(\tau(Q^l(K^l)^\top)V^l) + E^{l-1}) \quad (5)$$

where  $l \in \{1, 2, \dots, L\}$  is the number of encoder block,  $\sigma_Q^l$ ,  $\sigma_K^l$ ,  $\sigma_V^l$ , and  $\sigma^l$  are fully connected layers within the  $l$ -th encoder block. The term  $\tau$  refers to the SiLU activation, while  $\omega^l$  indicates layer normalization in the  $l$ -th block.

The fine-grained user interest representation yielded by the final encoder block is then fed into the set-wise decoder to model user-item interactions.

**4.3.2 Set-wise Decoder.** The set-wise decoder is composed of a series of decoder blocks. Each block contains two main components: a cross-item interaction block, which employs the self-attention mechanism to identify dependencies among all items within the current request, and a user-item interaction block, which utilizes the cross-attention mechanism to simultaneously capture the user's dynamic interests across all items. The input to the set-wise decoder is structured as follows: Initially, each item's profile and user-item

cross features are processed through a shared tokenization layer and projected into a  $d$ -dimensional item embedding. Subsequently, the user profile and context features are fed into another tokenization layer and projected into a  $d$ -dimensional user-context embedding. The item and user-context embeddings are then combined and fed into the set-wise decoder. This process can be formulated as:

$$H = \bar{\sigma}_1(\phi(f_u) || \phi(f_c)), \quad (6)$$

$$D_i^0 = \bar{\sigma}_2(\phi(f_i) || \phi(f_{u,i})) + H. \quad (7)$$

Here,  $H$  is the user-context embedding, and  $D_i^0$  denotes the input embedding of the set-wise decoder from the  $i$ -th item. The terms  $\bar{\sigma}_1$  and  $\bar{\sigma}_2$  are fully connected layers, while  $||$  indicates concatenation. Note that the input embeddings of the decoder share similar semantics as the initial user interest representation in the encoder.

Thereafter,  $D^0 = \{D_1^0, D_2^0, \dots, D_K^0\}$  is fed into a series of decoder blocks for modeling cross-item and user-item interactions, and the computations are conducted as follows:

$$\bar{Q}^m = \tau(\bar{\sigma}_Q^m(D^{m-1})), \bar{K}^m = \tau(\bar{\sigma}_K^m(D^{m-1})), \bar{V}^m = \tau(\bar{\sigma}_V^m(D^{m-1})), \quad (8)$$

$$\bar{D}^m = \bar{\omega}^m(\bar{\sigma}^m(\tau(\bar{Q}^m(\bar{K}^m)^\top)\bar{V}^m) + D^{m-1}), \quad (9)$$

$$\hat{Q}^m = \tau(\hat{\sigma}_Q^m(\bar{D}^m)), \hat{K}^m = \tau(\hat{\sigma}_K^m(E^L)), \hat{V}^m = \tau(\hat{\sigma}_V^m(E^L)), \quad (10)$$

$$\hat{D}^m = \hat{\omega}^m(\hat{\sigma}^m(\tau(\hat{Q}^m(\hat{K}^m)^\top)\hat{V}^m) + \bar{D}^m). \quad (11)$$

Here,  $m \in \{1, 2, \dots, M\}$  is the number of decoder blocks,  $\bar{\sigma}_Q^m$ ,  $\bar{\sigma}_K^m$ ,  $\bar{\sigma}_V^m$ ,  $\hat{\sigma}_Q^m$ ,  $\hat{\sigma}_K^m$ ,  $\hat{\sigma}_V^m$ , and  $\hat{\sigma}^m$  are fully connected layers within the  $m$ -th decoder block. The terms  $\bar{\omega}^m$  and  $\hat{\omega}^m$  indicate layer normalization in the  $m$ -th block.

Finally, the output yielded by the last decoder block is fed into two multilayer perceptrons (MLPs) to estimate the probability of an item being exposed to and clicked by the user:

$$p_i^{exp} = \text{sigmoid}(\text{MLP}_1(D_i^M)), \quad (12)$$

$$p_i^{clk} = \text{sigmoid}(\text{MLP}_2(D_i^M)). \quad (13)$$

HoMer integrates both sequential and set-wise contexts, significantly enhancing prediction performance. It shares user-specific computations across all items, and predicts for them in parallel, showcasing exceptional efficiency. Besides, HoMer decouples user-specific and item-specific computations, allowing each to be independently scaled up to further improve prediction performance.

#### 4.4 Training and Deployment

**4.4.1 Training Objective.** Traditional CTR prediction models typically utilize cross-entropy loss on the exposed items for training, which can be formulated as:

$$\mathcal{L}_{clk} = -\frac{1}{|\mathcal{I}^{exp}|} \sum_{i \in \mathcal{I}^{exp}} \left( y_i^{clk} \log p_i^{clk} + (1 - y_i^{clk}) \log(1 - p_i^{clk}) \right). \quad (14)$$

Here,  $\mathcal{I}^{exp}$  indicates all items exposed to users, and  $y_i^{clk} \in \{0, 1\}$  denotes whether the exposed item is clicked or not.

However, we have observed that relying solely on the loss function defined above may not provide sufficient signals for HoMer to effectively learn complex cross-item interactions. To address this issue, we apply an auxiliary impression loss, formulated as:

$$\mathcal{L}_{imp} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left( y_i^{exp} \log p_i^{exp} + (1 - y_i^{exp}) \log(1 - p_i^{exp}) \right). \quad (15)$$

Here,  $\mathcal{I}$  denotes all items generated by the pre-ranking stage, regardless of whether they are exposed to the user. The term  $y_i^{exp} \in \{0, 1\}$  indicates whether the item is exposed to the user. The final loss combines the above objectives with a balancing hyperparameter  $\lambda$ , which is set to 1 in this work:

$$\mathcal{L} = \mathcal{L}_{clk} + \lambda \mathcal{L}_{imp}. \quad (16)$$

We use the Adam optimizer [13] with a learning rate set to  $1e^{-4}$  for model training. Following industry best practices, we employ a one-epoch training strategy, meaning each sample is processed only once to prevent overfitting.

**4.4.2 Deployment.** In industrial recommender systems, the pre-ranking stage typically generates tens to hundreds of items. Traditional online CTR prediction services divide these items into multiple shards and perform model predictions for these shards in parallel to improve prediction efficiency. In our scenario, the 99th percentile of item count does not exceed 300. By configuring a large shard size, such as 300, HoMer can predict all items in approximately one model invocation. When the item count exceeds 300, HoMer must still partition them into several shards. Although this may prevent HoMer from fully capturing the entire item set context online, it remains effective as it has already learned cross-item interactions from requests containing fewer than 300 items.

As previously noted, the item count in the majority of requests does not exceed 300. We utilize Flash Attention [5] and jagged tensor within our attention-based modules to efficiently manage varying input lengths of the sequential and set-wise contexts. This approach eliminates redundant computations at padding positions during both the training and serving phases. Furthermore, due to the structural simplification and homogeneity inherent in HoMer, engineering optimizations applied to the encoder or decoder blocks can effectively enhance the overall efficiency of the model. By executing kernel fusion on these fundamental blocks, we significantly

improve Model FLOPs Utilization (MFU) from 7.8% to 12.2%. Coupled with HoMer's computational efficiency, we achieve a 27% reduction in online GPU resource consumption.

## 5 Experiments

In this section, we evaluate our proposed approaches on industrial dataset and aim to answer the following research questions:

- **RQ1:** How does our HoMer model perform in comparison to state-of-the-art DLRMs?
- **RQ2:** What is the improvement in model performance resulting from panoramic sequence modeling and cross-item interaction modeling?
- **RQ3:** How do hyperparameters affect model performance?
- **RQ4:** How does HoMer perform in a real recommender system?

### 5.1 Experimental Setup

**5.1.1 Dataset.** Our experiments are based on real interaction logs gathered from the search advertising scenario of Meituan. For offline experiments, we sample logs from April 2025 to July 2025 to build both point-wise and set-wise datasets. These datasets include 420 million requests from more than 39 million users, featuring nearly 690 thousands unique items and 1.25 billion user behaviors.

**5.1.2 Metrics.** We use **Log Loss** and Area Under Curve (AUC) as our offline evaluation metrics. Log Loss is the higher the better and AUC is the lower the better. Note that a 0.001 improvement in AUC is considered deployment-worthy in our system. For the online A/B testing, we use **CTR** and **RPM** (Revenue Per Mille) to measure improvement. Additionally, we examine the **dense parameter** of the model as well as the **FLOPs** (Floating Point Operations) to align the model's performance under consistent standards. For fairness, We calculate the **GFLOPs Per Request** for each model. Specifically, each request consists of up to 300 items.

**5.1.3 Baselines.** We implement six commonly used point-wise architectures considering different technical dimensions in industrial recommender systems to compare with HoMer. To ensure fair comparisons, the panoramic sequence is utilized across all models.

- **Point-wise Baseline.** As shown in Figure 1(c), this baseline involves sequence modeling components (e.g., DIN [45]), feature interaction components (e.g., DCN [28]), and feature enhancement components (e.g., MoE [23]).
- **SASRec** [12]. This baseline leverages the self-attention mechanism to model sequential dependencies in behavior sequence.
- **DCNv2** [29]. This baseline improves feature interaction learning in recommender systems through enhanced cross-network architecture and low-rank techniques.
- **Wukong** [36]. This method introduces a scalable architecture using stacked factorization machines, establishing reliable scaling laws for model performance as complexity increases.
- **CIM** [40]. This method introduces the auxiliary item set features and an impression submodule to improve point-wise prediction.
- **Point-wise HoMer.** A variant of HoMer that removes the cross-item interaction blocks and  $\mathcal{L}_{imp}$  from the set-wise decoder. In this model, although all items are predicted in one model invocation, the items cannot perceive each other. To ensure a fair comparison with HoMer, we increase the number of encoder

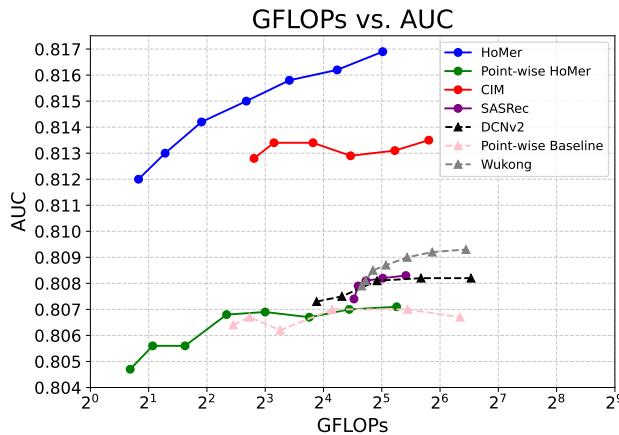
blocks and user-item interaction blocks to maintain a similar FLOPs.

## 5.2 Overall Performance (RQ1)

As shown in Table 1 and Figure 4, the proposed HoMer substantially outperforms all competing methods and exhibits a distinct scaling phenomenon. Compared to other heterogeneous models, HoMer demonstrates a more pronounced performance improvement as FLOPs increases. We attribute these advantages to the synergistic effects of feature, context and architecture heterogeneities. Specifically, the performance of point-wise HoMer improves with increasing FLOPs at lower computational budgets, but there exists a threshold beyond which additional scaling yields diminishing returns. The saturation suggests that it is more beneficial to allocate additional FLOPs to address context heterogeneity. Moreover, while CIM's base AUC is notably higher due to the modeling of cross-item interaction context, its performance fails to improve with increased model complexity. We attribute this to the see-saw effect between different network modules introduced by architecture heterogeneity.

**Table 1: The main experimental results of HoMer compared with other six baselines. The notation  $\uparrow$  indicates the higher the better, and the notation  $\downarrow$  denotes the lower the better.**

Models	AUC $\uparrow$	LogLoss $\downarrow$	Flops	Params
Point-wise Baseline	0.8070	0.2464	46.8G	71.8M
DCNv2	0.8082	0.2458	99.3G	159.1M
Wukong	0.8093	0.2442	93.5G	120.6M
SASRec	0.8081	0.2460	45.7G	58.6M
CIM	0.8134	0.2438	60.1G	67.4M
Point-wise HoMer	0.8071	0.2464	40.9G	92.3M
HoMer	0.8169	0.2425	42.6G	114.9M



**Figure 4: Scalability of models with respect to FLOPs.**

## 5.3 Ablation Study (RQ2)

**5.3.1 Effects of Panoramic Sequence Modeling.** In Section 4.1, we propose a panoramic sequence by aligning sequence side features with non-sequential features. The side features are categorized into four domains, including user/item profiles, user-item cross features, and context features. We conduct an ablation study to assess the effects of these side feature domains, with the experimental results encapsulated in Table 2. The analysis reveals that each domain contributes positively to HoMer's performance, and their synergistic integration markedly enhances the AUC metric from 0.8128 to 0.8169. This underscores the notion that these side features harbor complementary information, warranting their comprehensive utilization.

**Table 2: Ablation of side features in panoramic sequence.**

User	Item	User-Item	Context	AUC $\uparrow$	LogLoss $\downarrow$	side info %
				0.8128	0.2440	0.0%
✓				0.8149	0.2434	14.8%
	✓			0.8154	0.2432	36.3%
		✓		0.8152	0.2431	37.9%
			✓	0.8150	0.2433	11.0%
✓	✓	✓	✓	0.8169	0.2425	100.0%

**5.3.2 Effects of cross-item interaction modeling.** In Section 4.3.2 and Section 4.4.1, we delineate the cross-item interaction block and an auxiliary impression loss  $\mathcal{L}_{imp}$ , aimed at modeling the contextual relationships among items and capturing authentic user behavior patterns. To evaluate the performance enhancements attributable to these designs, we perform an ablation study on HoMer and its three variants: point-wise HoMer, HoMer excluding impression loss (HoMer w/o ImpLoss), and HoMer devoid of cross-item interaction blocks (HoMer w/o cross-item). Furthermore, we assess the benefits of these designs across varying model sizes by adjusting the number of transformer blocks. For equitable comparison, we elucidate the correlation between the number of dense parameters in the model and the AUC metric. As depicted in Figure 5, the comparison between the green and red lines with the purple line demonstrates that integrating either the cross-item interaction block or the auxiliary impression loss results in substantial improvements. The blue and red lines reveal that the inclusion of cross-item interaction context significantly enhances AUC as the model size increases. A comparison of the blue and green lines indicates that the auxiliary impression loss furnishes effective signals, enabling HoMer to more adeptly learn cross-item interaction context.

## 5.4 Hyperparameter Analysis (RQ3)

We observe that increasing the depth of model, as well as expanding the dimensionality of embeddings and tokens, consistently leads to significant performance improvements.

**5.4.1 Depth of HoMer.** Fig. 6 illustrates the effects of varying block configurations within the sequential encoder and set-wise decoder. The blue bars represent configurations with a constant decoder depth while varying the encoder depth, indicating that augmenting solely the number of encoder blocks results in only slight improvements. Conversely, the orange, green, and red bars reveal that both

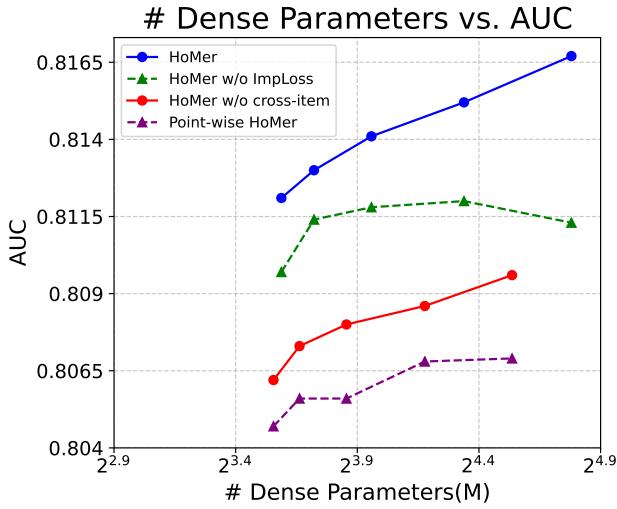


Figure 5: Ablation of cross-item interaction modeling.

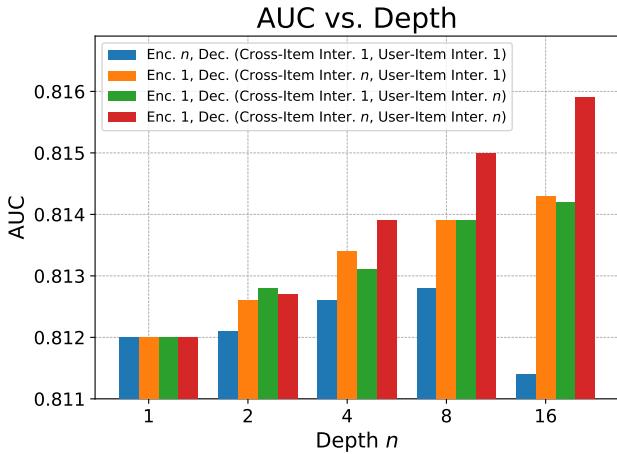


Figure 6: The effects of the depth of HoMer.

*cross-item interaction blocks* and *user-item interaction blocks* contribute comparably to performance enhancements, and these two blocks complement each other effectively. These observations imply that HoMer’s performance exhibits greater sensitivity to the capacity of the set-wise decoder.

**5.4.2 Dimensionality of embeddings and tokens.** As presented in Table 3, increasing the token dimension ( $D_{token}$ ) consistently enhances the model’s predictive performance, as reflected by higher AUC values across various layer configurations. For example, when the token dimension is doubled from 256 to 512, the AUC improves from 0.8150 to 0.8162 for  $L = 8, M = 8$ , and from 0.8161 to 0.8167 for  $L = 8, M = 16$ . This trend indicates that a larger token dimension enables the model to capture richer and more informative representations, thereby boosting its discriminative capability. On the other hand, increasing the embedding dimension ( $D_{embed}$ ) from

Table 3: Prediction performance and computational burden of HoMer with varying model depths, embedding dimensions, and token dimensions.

Layers #	$D_{embed}$	$D_{token}$	AUC ↑	Flops	Params
$L = 8, M = 8$	16	256	0.8150	6.9G	21.2M
	32	256	0.8151	8.9G	35.6M
	16	512	0.8162	20.2G	55.8M
	32	512	0.8164	24.3G	84.5M
$L = 8, M = 16$	16	256	0.8161	11.4G	28.8M
	32	256	0.8163	15.4G	46.3M
	16	512	0.8167	34.8G	79.9M
	32	512	0.8169	42.6G	114.9M

16 to 32 results in only slight improvements in AUC, suggesting that the model’s performance is less sensitive to changes in embedding dimensionality compared to token dimensionality. These observations imply that allocating more computational resources to expanding the token dimension is generally more beneficial for improving model accuracy, while increasing the embedding dimension yields diminishing returns. However, it is important to note that higher token and embedding dimensions also lead to increased computational burden, as indicated by the substantial growth in FLOPs and parameter counts. Therefore, a careful trade-off between performance and efficiency should be considered when selecting model dimensions.

### 5.5 Online A/B Test (RQ4)

We conduct A/B testing on Meituan’s search advertising scenario with 20% online traffic between 2025-09-08 and 2025-09-15. HoMer achieved gains of CTR+1.99% and RPM+2.46% over the highly optimized baseline, validating its superior effectiveness. Now, HoMer has been deployed on the main traffic of Meituan’s search advertising, serving tens of millions of users.

## 6 Conclusion

This study identifies three types of heterogeneities that undermine CTR prediction performance: feature, context, and architecture heterogeneities. We introduce HoMer, a Homogeneous-Oriented Transformer, to mitigate the aforementioned heterogeneities. HoMer employs a unified encoder-decoder architecture, with the sequential encoder dedicated to capturing fine-grained user interest representation from the proposed panoramic sequence, and the set-wise decoder focused on modeling cross-item interaction context and learning authentic user behavior patterns. Comprehensive experiments conducted in Meituan’s search advertising scenario demonstrate HoMer’s superiority, computational efficiency and scalability. The online A/B test further underscore HoMer’s practical value for large-scale recommender systems. In future work, we aim to explore engineering strategies to enhance HoMer’s scalability, including more efficient training and inference techniques, as well as system-level optimizations, to fully realize its potential in even larger and more complex industrial scenarios.

## References

- [1] Jianxin Chang, Chenbin Zhang, Zhiyi Fu, Xiaoxue Zang, Lin Guan, Jing Lu, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, et al. 2023. 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*. 3785–3794.
- [2] Chi Chen, Hui Chen, Kangzhi Zhao, Junsheng Zhou, Li He, Hongbo Deng, Jian Xu, Bo Zheng, Yong Zhang, and Chunxiao Xing. 2022. Extr: click-through rate prediction with externalities in e-commerce sponsored search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2732–2740.
- [3] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.
- [4] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.
- [5] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems* 35 (2022), 16344–16359.
- [6] Yufei Feng, Fuyu Lv, Weichen Shen, Menghan Wang, Fei Sun, Yu Zhu, and Keping Yang. 2019. Deep session interest network for click-through rate prediction. *arXiv preprint arXiv:1905.06482* (2019).
- [7] Huifeng Guo, Ruiming Tang, Yuming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1725–1731.
- [8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.
- [9] Yupeng Hou, Jianmo Ni, Zankui He, Naveen Sachdeva, Wang-Cheng Kang, Ed H. Chi, Julian McAuley, and Derek Zhiyuan Cheng. 2025. ActionPiece: Contextually Tokenizing Action Sequences for Generative Recommendation. doi:10.48550/arXiv.2502.13581 arXiv:2502.13581 [cs].
- [10] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
- [11] Tongwen Huang, Zhiqi Zhang, and Junlin Zhang. 2019. FiBiNET: combining feature importance and bilinear feature interaction for click-through rate prediction. In *Proceedings of the 13th ACM conference on recommender systems*. 169–177.
- [12] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.
- [13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [14] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.
- [15] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1754–1763.
- [16] Chang Liu, Xiaoguang Li, Guohao Cai, Zhenhua Dong, Hong Zhu, and Lifeng Shang. 2021. Noninvasive self-attention for side information fusion in sequential recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4249–4256.
- [17] Enze Liu, Bowen Zheng, Cheng Ling, Lantao Hu, Han Li, and Wayne Xin Zhao. 2025. Generative Recommender with End-to-End Learnable Item Tokenization. doi:10.48550/arXiv.2409.05546 arXiv:2409.05546 [cs].
- [18] Zihan Liu, Yupeng Hou, and Julian McAuley. 2024. Multi-Behavior Generative Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 1575–1585. doi:10.1145/3627673.3679730
- [19] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. 2019. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091* (2019).
- [20] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. 2019. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091* (2019).
- [21] Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. 2020. 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*. 2685–2692.
- [22] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems* 36 (2023), 10299–10315.
- [23] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).
- [24] Zihua Si, Lin Guan, ZhongXiang Sun, Xiaoxue Zang, Jing Lu, Yiqun Hui, Xingchao Cao, Zeyu Yang, Yichen Zheng, Dewei Leng, et al. 2024. 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*. 4890–4897.
- [25] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1161–1170.
- [26] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
- [27] Chunqi Wang, Bingchao Wu, Zheng Chen, Lei Shen, Bing Wang, and Xiaoyi Zeng. 2025. Scaling Transformers for Discriminative Recommendation via Generative Pretraining. *arXiv preprint arXiv:2506.03699* (2025).
- [28] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*. 1–7.
- [29] Ruoxi Wang, Rakesh Shrivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*. 1785–1797.
- [30] Yifan Wang, Weinan Gan, Longtao Xiao, Jieming Zhu, Heng Chang, Haozhao Wang, Rui Zhang, Zhenhua Dong, Ruiming Tang, and Ruixuan Li. 2025. Act-With-Think: Chunk Auto-Regressive Modeling for Generative Recommendation. doi:10.48550/arXiv.2506.23643 arXiv:2506.23643 [cs].
- [31] Ye Wang, Jiahao Xun, Minjie Hong, Jieming Zhu, Tao Jin, Wang Lin, Haoyuan Li, Linjun Li, Yan Xia, Zhou Zhao, and Zhenhua Dong. 2024. EAGER: Two-Stream Generative Recommender with Behavior-Semantic Collaboration. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. Association for Computing Machinery, New York, NY, USA, 3245–3254. doi:10.1145/3637528.3671775 event-place: Barcelona, Spain.
- [32] Yueqi Xie, Peilin Zhou, and Sungjun Kim. 2022. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1611–1621.
- [33] Chaoqun Yang, Xinyu Lin, Wenjie Wang, Yongqi Li, Teng Sun, Xianjing Han, and Tat-Seng Chua. 2025. EARN: Efficient Inference Acceleration for LLM-based Generative Recommendation by Register Tokens. doi:10.48550/arXiv.2507.00715 arXiv:2507.00715 [cs].
- [34] Yuhao Yang, Zhi Ji, Zhaopeng Li, Yi Li, Zhonglin Mo, Yue Ding, Kai Chen, Zijian Zhang, Jie Li, Shuanglong Li, and Lin Liu. 2025. Sparse Meets Dense: Unified Generative Recommendations with Cascaded Sparse-Dense Representations. doi:10.48550/arXiv.2503.02453 arXiv:2503.02453 [cs].
- [35] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, Yinghai Lu, and Yu Shi. 2024. Actions speak louder than words: trillion-parameter sequential transducers for generative recommendations. In *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*, Vol. 235. JMLR.org, Vienna, Austria, 58484–58509.
- [36] Buyun Zhang, Liang Luo, Yuxin Chen, Jade Nie, Xi Liu, Shen Li, Yanli Zhao, Yuchen Hao, Yantao Yao, Ellis Dingqiao Wen, et al. 2024. Wukong: towards a scaling law for large-scale recommendation. In *Proceedings of the 41st International Conference on Machine Learning*. 59421–59434.
- [37] Luankang Zhang, Kenan Song, Yi Quan Lee, Wei Guo, Hao Wang, Yawen Li, Huifang Guo, Yong Liu, Defu Lian, and Enhong Chen. 2025. Killing Two Birds with One Stone: Unifying Retrieval and Ranking with a Single Generative Recommendation Model. doi:10.48550/arXiv.2504.16454 arXiv:2504.16454 [cs].
- [38] Tingting Zhang, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Deqing Wang, Guanfeng Liu, Xiaofang Zhou, et al. 2019. Feature-level deeper self-attention network for sequential recommendation.. In *IJCAI*. 4320–4326.
- [39] Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting Large Language Models by Integrating Collaborative Semantics for Recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. 1435–1448. doi:10.1109/ICDE60146.2024.00118 ISSN: 2375-026X.
- [40] Kaifu Zheng, Lu Wang, Yu Li, Xusong Chen, Hu Liu, Jing Lu, Xiwei Zhao, Changping Peng, Zhangang Lin, and Jingping Shao. 2022. Implicit user awareness modeling via candidate items for ctr prediction in search ads. In *Proceedings of the ACM Web Conference 2022*. 246–255.
- [41] Guorui Zhou, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Shiyao Wang, Weifeng Ding, Wuchao Li, Xincheng Luo, Xingmei Wang, Zexuan Cheng, Zixing Zhang, Bin Zhang, Boxuan Wang, Chaoyi Ma, Chengru Song, Chenhui Wang, Di Wang, Dongxue Meng, Fan

- Yang, Fangyu Zhang, Feng Jiang, Fuxing Zhang, Gang Wang, Guowang Zhang, Han Li, Hengrui Hu, Hezheng Lin, Hongtao Cheng, Hongyang Cao, Huanjie Wang, Jiaming Huang, Jiapeng Chen, Jiaqiang Liu, Jinghui Jia, Kun Gai, Lantao Hu, Liang Zeng, Liao Yu, Qiang Wang, Qidong Zhou, Shengzhe Wang, Shihui He, Shuang Yang, Shujie Yang, Sui Huang, Tao Wu, Tiantian He, Tingting Gao, Wei Yuan, Xiao Liang, Xiaoxiao Xu, Xugang Liu, Yan Wang, Yi Wang, Yiwu Liu, Yue Song, Yufei Zhang, Yunfan Wu, Yunfeng Zhao, and Zhanyu Liu. 2025. OneRec Technical Report. doi:10.48550/arXiv.2506.13695 arXiv:2506.13695 [cs].
- [42] Guorui Zhou, Hengrui Hu, Hongtao Cheng, Huanjie Wang, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Lu Ren, Liao Yu, et al. 2025. OneRec-V2 Technical Report. *arXiv preprint arXiv:2508.20900* (2025).
- [43] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.
- [44] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.
- [45] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.
- [46] Jie Zhu, Zhifang Fan, Xiaoxie Zhu, Yuchen Jiang, Hangyu Wang, Xintian Han, Haoran Ding, Xinmin Wang, Wenlin Zhao, Zhen Gong, et al. 2025. RankMixer: Scaling Up Ranking Models in Industrial Recommenders. *arXiv preprint arXiv:2507.15551* (2025).