

SpecFormer: Mitigating Embedding and Attention Collapse via Spectral-Aware Transformer for Recommendation

1st Yu Cui *
Zhejiang University
Hangzhou, China
cuiyu23@zju.edu.cn
ORCID: 0009-0001-6203-3022

2nd Yi Xu*
Alibaba Group
Beijing, China
xy397404@alibaba-inc.com
ORCID:0009-0007-3571-8791

3rd Jiahao Wang
Alibaba Group
Beijing, China
wjh177423@alibaba-inc.com
ORCID:0009-0000-3046-2351

4rd Hao Zhang
Alibaba Group
Beijing, China
zh138764@alibaba-inc.com
ORCID:0009-0002-3137-4980

5rd Yu Zhang
Alibaba Group
Beijing, China
daoji@alibaba-inc.com
ORCID:0000-0002-6057-7886

6rd Xiaoyi Zeng
Alibaba Group
Beijing, China
yuanhan@taobao.com
ORCID:0000-0002-3742-4910

7th Can Wang
Zhejiang University
Hangzhou, China
wcan@zju.edu.cn
ORCID: 0000-0002-5890-4307

8rd Jinxin Hu[†]
Alibaba Group
Beijing, China
jinxin.hjx@alibaba-inc.com
ORCID:0000-0002-7252-5207

9th Jiawei Chen[†]
Zhejiang University
Hangzhou, China
sleepyhunt@zju.edu.cn
ORCID: 0000-0002-4752-2629

Abstract—Transformer architectures have achieved remarkable success across diverse domains; however, directly applying their standard self-attention mechanism to recommendation often yields suboptimal performance, sometimes even trailing behind well-designed simple recommendation models. In this paper, we reveal that this performance bottleneck stems from severe embedding and attention collapse unique to recommendation scenarios. The heterogeneity and long-tail nature of recommendation data lead to a severe spectral collapse dominated by a few principal singular values. We further theoretically demonstrate that this triggers a vicious cycle in recommendation model’s forward and backward propagation, which accelerates embedding and attention collapse and limits the model’s scaling capability with increased depth.

To address these issues, we propose SpecFormer, a novel Spectral-Aware Transformer designed for mitigating embedding and attention collapse in recommendation. Specifically, SpecFormer introduces 1) a Learnable Spectral Softening module to dynamically smooth the singular values distribution of the input token embeddings; 2) a Spectrum-softened Attention mechanism to model feature interaction under a more uniform spectral distribution space; 3) a Spectral Residual Position Encoding via Taylor expansion of singular values, explicitly providing a spectral inductive bias for feature interactions. Extensive experiments on one industrial and two public datasets demonstrate that SpecFormer significantly outperforms state-of-the-art baselines. Notably, SpecFormer has been successfully deployed in a real-world commercial recommender system and exhibits exceptional scaling capabilities: stacking SpecFormer layers ac-

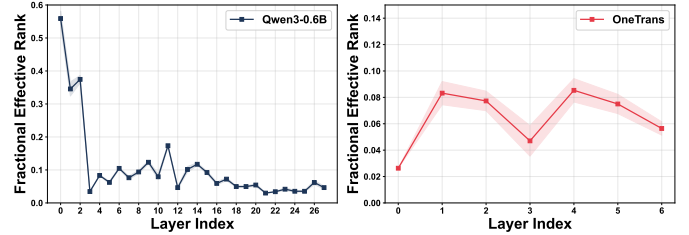


Fig. 1: The comparison on fractional effective rank of attention matrix with the increase of Transformer layers. Unlike LLMs, attention matrix in recommendation suffers from serious collapse at the early Transformer layers.

tively improves the attention effective rank and recommendation performance.

Index Terms—Recommender System, Transformer Architecture, Attention Mechanism, Spectral Collapse

I. INTRODUCTION

Transformer architectures [1] have achieved remarkable success across Natural Language Processing (NLP) [2], [3] and Computer Vision (CV) [4], [5], largely attributed to the expressive power of the self-attention mechanism in capturing long-range dependencies. This success has motivated extensive efforts to adapt Transformers for Recommender Systems (RS) [6], [7]. Beyond their initial applications in sequen-

*Equal contribution. [†]Corresponding authors.

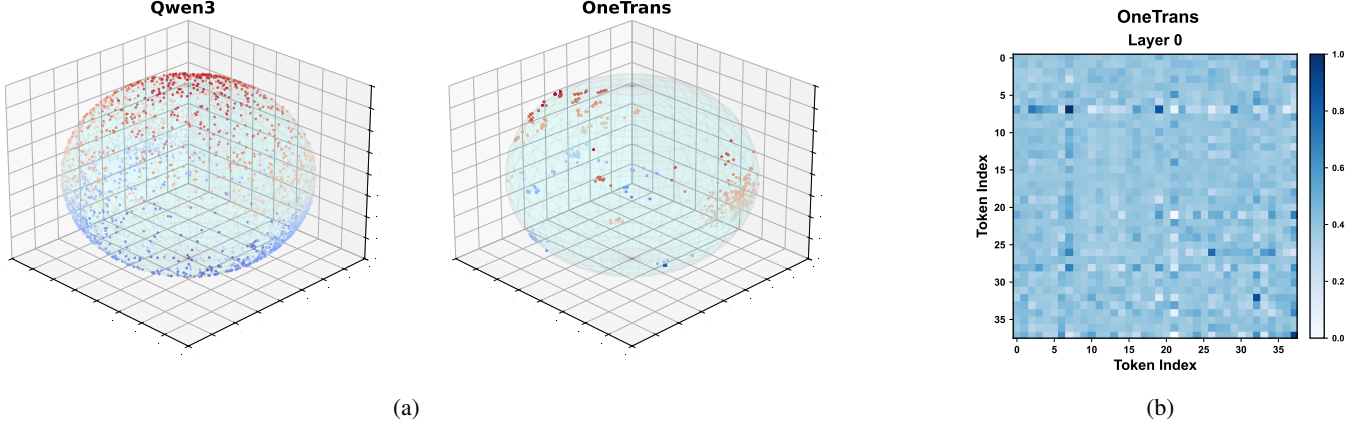


Fig. 2: (a) Visualization of the token semantic space distributions of Qwen3 and OneTrans, indicating evident heterogeneity in the recommendation semantic space; (b) Heatmap of the first-layer attention score matrix of OneTrans, displaying a highly uniform distribution. A similar phenomenon is observed across various data samples.

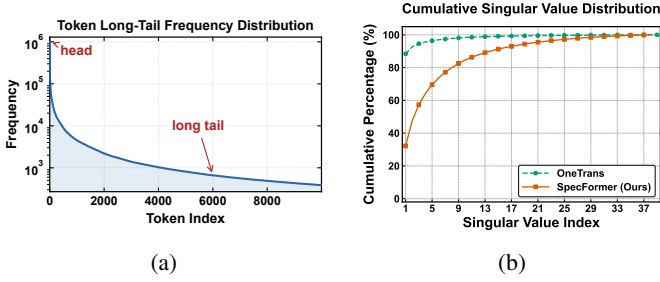


Fig. 3: (a) Token frequency exhibits a severe long-tail distribution in recommendation. (b) Cumulative singular value distribution of OneTrans and our method. OneTrans suffers from severe embedding collapse where top singular values dominate, and SpecFormer effectively mitigates this collapse.

tial recommendation [8], [9], Transformers have increasingly emerged as the core architecture for feature interaction modeling. By regarding diverse feature data as individual tokens, the attention mechanism can extract high-order cross-feature dependencies, which has been widely deployed in large-scale industrial recommendation applications [10], [11].

However, directly applying the standard self-attention mechanism to recommendation scenarios often hits a severe performance bottleneck. Our empirical observations (*cf.* Table III) reveal that the state-of-the-art attention-based feature interaction models (*e.g.*, OneTrans [6]) sometimes even trail behind carefully designed simple feature interaction architectures (*e.g.*, RankMixer [11]). More critically, unlike LLMs, which benefit significantly from depth [12], Transformers in recommendation suffer from serious attention collapse at the early Transformer layers (*cf.* Figure 1), which results in poor scaling capabilities: stacking layers often leads to severe performance degradation and a sharp drop in the effective rank of the attention matrix (*cf.* Section V-D). This naturally raises a key question: **Why does the attention mechanism**

yield sub-optimal performance and collapse so quickly in recommender systems?

Through empirical and theoretical analysis, we reveal that it is the heterogeneity and long-tail distribution characteristics of recommendation data that catalyzes a vicious cycle of spectral collapse for both feature embeddings and attention maps in recommendation:

(1) Inherent Data Heterogeneity and Long-Tail Distributions Induce Spectral Collapse. Unlike LLMs, where text tokens share a relatively continuous and smooth semantic space, recommendation features (*e.g.*, user profiles, item IDs, and contexts) are highly heterogeneous. Our empirical analysis illustrates that recommendation token embeddings do not distribute uniformly; rather, they form highly segregated, discrete semantic sub-spaces (*cf.* Figure 2a). Consequently, when attention mechanisms attempt to model these tokens, they fail to establish smooth semantic alignments across these heterogeneous fields, leading to overly sparse attention maps (*cf.* Figure 2b). Furthermore, recommendation data inherently exhibits a severe long-tail distribution in feature frequencies, which leads to the long-tail distribution of input feature tokens (*cf.* Figure 3a). As demonstrated by previous work [13], this triggers a critical *spectral collapse* in the representation space, where the embedding matrix is dominated by a few principal singular values, while the minor spectral components are essentially obliterated (*cf.* Figure 3b).

(2) The Vicious Cycle of Collapse in Forward and Backward Propagation. Most alarmingly, we theoretically demonstrate that this spectral collapse is not merely a static flaw, but triggers a vicious cycle within the Transformer architecture itself (*cf.* Section III). In the forward propagation, due to the collapsed input token embeddings, the computed attention matrix degenerates into a sparse, low-rank form dominated almost entirely by the principal spectral components. As layers stack, this low-rank attention acts as a low-pass filter, aggressively smoothing out token representations.

During backward propagation, gradients flowing through this low-rank attention matrix are forcefully projected onto the dominant principal sub-spaces. Consequently, minor spectral directions suffer from gradient starvation, receiving almost no effective updates. This biased update rule further aggravates the spectral collapse in the next iteration. This vicious cycle explains why Transformers in recommendation collapse at much shallower layers compared to the LLMs (*cf.* Figure 1).

While several existing efforts have explored various techniques to mitigate the spectral collapse in recommendation, they still face critical limitations. Specifically, some methods attempt to introduce multi-embedding paradigms [14], [15]; however, this introduces prohibitive parameter overhead, making it highly impractical for large-scale systems. Other approaches introduce extensive FFN parameters [11] or external knowledge integration [16] to avoid embedding collapse but merely operate in the original spatial domain. However, they fail to fundamentally resolve the root cause of spectral collapse, which yield sub-optimal results (*cf.* Table III for empirical validation). While a few recent works have explored spectral regularizations [17] or Fourier transforms [18], they are either unsuitable for feature interaction modeling or fail to improve the attention collapse. Consequently, exploring a novel Transformer architecture that intrinsically resolves the collapse bottleneck directly from a spectral perspective is of paramount importance.

To tackle these fundamental challenges, we propose **SpecFormer**, a novel **Spectral-Aware Transformer** architecture specifically tailored for recommendation. SpecFormer breaks the vicious cycle by operating feature interactions explicitly within a dynamically softened spectral domain. Specifically, 1) we first introduce a **Learnable Spectral Softening** module that dynamically re-weights the singular values of the input representations, yielding a more uniform spectral distribution that suppresses dominant components and revitalizes tail features. 2) Building upon this, we propose a **Spectrum-softened Attention** mechanism that computes attention weights within this softened spectral space, forcing the model to capture spectral-aware feature interaction signals. 3) Finally, to provide explicit spectral inductive biases, we design a **Spectral Residual Position Encoding** derived from the Taylor expansion-based mapping function of singular values for attention bias.

Extensive experiments on one industrial and two public benchmark datasets demonstrate the superiority of SpecFormer. It not only significantly outperforms state-of-the-art baselines but also exhibits exceptional scaling capabilities: unlike standard Transformers, stacking layers in SpecFormer actively improves both recommendation performance and the effective rank of attention matrices. **Notably, SpecFormer has been successfully deployed in a real-world e-commerce advertising platform. In a large-scale online A/B test, it yielded tremendous business gains: a +1.34% increase in CTR and a +16.72% boost in orders at an acceptable cost of only a 5ms latency overhead.**

Our main contributions are summarized as follows:

- We systematically uncover the root cause of Transformer’s

TABLE I: Notations in this paper.

Notation	Description
$\mathcal{F}$	Fields of recommendation data
L	Number of feature tokens
D	Token embedding dimension size
$\mathbf{X}$	Initial input feature token matrix
$\mathbf{H}^{(l)}$	Hidden token representation at layer l
$\mathbf{Q}, \mathbf{K}, \mathbf{V}_{\text{val}}$	Query, Key, and Value matrices in attention
$\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$	Projection matrices for Query, Key and Value
$\mathbf{O}$	The output matrix of self-attention
$\mathbf{S}$	Pre-softmax attention score matrix
r	Rank of the matrix
$\mathbf{U}$	Left singular matrix after SVD
$\mathbf{V}$	Right singular matrix after SVD
Σ	Singular value matrix
σ_i	The i -th singular value
Σ^*	Softened singular matrix
τ, a	Softening factor and its learnable parameter
$\mathbf{H}^*$	Softened token feature matrix
$\mathbf{P}_{\text{bias}}$	Learnable spectral residual position encoding bias
$\mathbf{S}_{\text{bias}}$	Original-space residual bias
β_p	Taylor expansion coefficient for the p -th term
α, γ	Learnable parameters for residual bias strength

attention mechanism failure in recommendation. We theoretically and empirically demonstrate the vicious cycle of attention and embedding collapse caused by data heterogeneity and long-tail distributions.

- We propose SpecFormer, a spectral-aware Transformer that resolves the collapse bottleneck intrinsically from the attention mechanism via Learnable Spectral Softening, Spectrum-softened Attention, and Spectral Residual Position Encoding.
- Extensive offline experiments and online A/B testing verify that SpecFormer achieves state-of-the-art performance. Crucially, it actively increases the attention’s effective rank and unlocks the depth-scaling potential of Transformers for recommendation.

II. PRELIMINARY

A. Task Formulation

This paper focuses on Click-Through Rate (CTR) prediction, a crucial and fundamental task in recommender systems [19]–[22]. It aims to estimate the probability that a user will click on a given item, based on a diverse set of categorical and numerical input features [23], [24].

Let $\mathcal{F} = \{f_1, f_2, \dots, f_F\}$ denote a set of F raw input fields (*e.g.*, user ID, age, item ID, temporal sequences). The CTR prediction task can be formulated as learning a predictive model f_θ :

$$\hat{y} = \sigma(f_\theta(\mathcal{F})), \quad (1)$$

where $\hat{y}$ is the predicted click probability and $\sigma(\cdot)$ is the sigmoid function. The core of f_θ lies in capturing effective feature interactions. Over the years, a variety of architectures have been proposed for f_θ , ranging from traditional FM-based models [25], [26] and DNN-based models [23], [27] to modern Transformer-based models [10], [28]. Among these, Transformer-based architectures have emerged as the

mainstream paradigm due to their exceptional capability in modeling complex, high-order feature interactions [6], [11].

B. Transformer-Based Feature Interaction Model

A standard Transformer-based feature interaction model typically consists of three sequential stages: input feature tokenization, feature interaction network, and feed-forward network for final prediction.

1) Input Feature Tokenization. Traditionally, each raw feature field is mapped to a distinct embedding and fed directly into the model [10]. However, treating hundreds of raw features as individual inputs inevitably leads to parameter fragmentation, insufficient modeling of important features, and under-utilization of GPU computation. To address this, modern Transformer-based models adopt a feature tokenization approach [6], [11]. Specifically, the features are first grouped into semantically coherent clusters. The embeddings of these grouped features are sequentially concatenated into a single unified vector $\mathbf{e}_{\text{input}} = [\mathbf{e}_1; \mathbf{e}_2; \dots; \mathbf{e}_L]$. Subsequently, $\mathbf{e}_{\text{input}}$ is partitioned and projected into an appropriate number of tokens with fixed dimensions:

$$\mathbf{x}_i = \text{Proj}(\mathbf{e}_{\text{input}}[d \cdot (i-1) : d \cdot i]), \quad i = 1, \dots, L, \quad (2)$$

where d is the dimension size of each partitioned chunk, L is the resulting token count, and the $\text{Proj}(\cdot)$ is the MLP that maps the split embedding into a unified dimension D . The tokenized feature matrix is then defined as $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_L] \in \mathbb{R}^{L \times D}$.

2) Feature Interaction Network. Taking the tokenized matrix $\mathbf{H}^{(0)} = \mathbf{X}$ as input, the model relies on feature interaction networks to model user preferences. The representation is updated via a residual connection and Layer Normalization (LN) [29]:

$$\mathbf{Z}^{(l)} = \text{LN}\left(\text{FeatureInteraction}\left(\mathbf{H}^{(l)}\right) + \mathbf{H}^{(l)}\right). \quad (3)$$

The $\text{FeatureInteraction}(\cdot)$ module explicitly models the cross-feature relations, and there are various designs for this component. A common approach is utilizing the standard self-attention mechanism (*e.g.*, OneTrans [6]). In contrast to standard attention, some recent works introduce carefully designed simplified mechanisms (*e.g.*, RankMixer [11]) to achieve efficient feature interaction without heavy projection matrices.

3) Feed-Forward Network. Following the interaction module, a Feed-Forward Network (FFN) is applied to further transform the token representations. While traditional Transformers share FFN parameters across all tokens, recent mainstream Transformer-based models often adopt a Per-token FFN (PFFN) [11] as an advanced alternative. By applying dedicated transformations to each token, PFFN preserves the diversity of disparate semantic sub-spaces and prevents high-frequency features from drowning out long-tail signals. The token updating process is formulated as:

$$\mathbf{H}^{(l+1)} = \text{LN}\left(\mathbf{Z}^{(l)} + \text{PFFN}\left(\mathbf{Z}^{(l)}\right)\right). \quad (4)$$

After propagating through several layers, the final representation is aggregated (*e.g.*, via mean pooling) and fed into a prediction head to yield the final predicted CTR $\hat{y}$.

C. Embedding Collapse and Attention Collapse

To formally quantify the collapse phenomena in Transformer-based recommendation, we first introduce the Singular Value Decomposition (SVD) [30]. For a matrix $\mathbf{H} \in \mathbb{R}^{L \times D}$ with rank $r = \min(L, D)$, the SVD can be expressed as:

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad (5)$$

where $\mathbf{U} \in \mathbb{R}^{L \times r}$ and $\mathbf{V} \in \mathbb{R}^{D \times r}$ are orthonormal matrices whose columns are the left and right singular vectors, respectively, and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0$ is the singular value matrix.

1) Embedding Collapse. Embedding collapse refers to the phenomenon where the input token representations degenerate into a narrow, low-dimensional subspace [31]. To evaluate this, we apply SVD on the token embedding matrix $\mathbf{H} \in \mathbb{R}^{L \times D}$. In the typical CTR setting, $L < D$, meaning the maximum possible rank is $r = L$. Following previous work [32], we use the *cumulative singular value ratio* to quantify the degree of embedding dimension collapse:

$$\rho(k) = \frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^r \sigma_i^2}, \quad k = 1, \dots, r. \quad (6)$$

A rapid saturation of $\rho(k)$ toward almost 100% at a small singular value index k indicates severe embedding collapse, meaning the feature space is dominated by only a few principal components.

2) Attention Collapse. Attention collapse occurs when the attention map loses its discriminative capability, degenerating into a smoothed or overly sparse pattern. To measure this, we apply SVD on the attention score matrix $\mathbf{S} \in \mathbb{R}^{L \times L}$ (where $r = L$). We quantify this degeneration using the *Effective Rank* [33], which calculates the Shannon entropy [34] of the singular value distribution. Effective Rank is also widely used in LLMs to evaluate attention collapse [1], [35], [36]. First, we compute the L_1 -normalized singular values p_i as:

$$p_i = \frac{\sigma_i}{\sum_{j=1}^r \sigma_j}, \quad i = 1, \dots, r, \quad (7)$$

where σ_i is the i -th singular value of $\mathbf{S}$. Then, the Effective Rank (erank) is defined as the exponential of the Shannon entropy of this distribution:

$$\text{erank}(\mathbf{S}) = \exp\left(-\sum_{i=1}^r p_i \ln p_i\right). \quad (8)$$

The value of $\text{erank}(\mathbf{S})$ ranges continuously from 1 (a completely collapsed rank-1 matrix) to r (a perfectly isotropic full-rank matrix). Furthermore, to provide a scale-invariant metric that facilitates fair comparisons across different sequence lengths or model dimensions, we use the *Fractional Effective*

Rank [15], [37] (ferank) by normalizing the Effective Rank by the maximum possible rank r :

$$\text{ferank}(\mathbf{S}) = \frac{\text{erank}(\mathbf{S})}{r}. \quad (9)$$

The $\text{ferank}(\mathbf{S})$ inherently bounds the metric within the range $[0, 1]$. A fractional effective rank approaching 0 serves as a standardized indicator of severe attention collapse, whereas a value of 1 signifies an optimal, perfectly full-rank attention map regardless of the absolute matrix size.

III. THEORETICAL ANALYSIS

In contrast to LLMs where token embeddings typically exhibit rich semantic variance, recommendation systems heavily rely on sparse recommendation data with severe power-law distributions (*cf.* Figure 2 and 3). This unique characteristic often leads to severe spectral collapse right at the input embedding layer (*cf.* Figure 1). In this section, we formally demonstrate that this inherent embedding collapse in recommendation scenarios triggers a vicious cycle of degradation through both the forward and backward propagation of standard attention mechanism. For mathematical simplicity, we omit the standard scaling factor $1/\sqrt{D}$.

Proposition III.1 (Forward Propagation: Severe Recommendation Embedding Collapse Induces Rapid Attention Collapse). *Let the input feature matrix be $\mathbf{H} \in \mathbb{R}^{L \times D}$ with its SVD denoted as $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_r)$. Given the severe spectral collapse premise in recommendation scenarios, assume $\sigma_k \leq \epsilon\sigma_1$ for all $k \geq 2$ ($0 < \epsilon \ll 1$). Then, the attention map degenerates and is almost entirely dominated by the principal spectral component, losing feature discriminability.*

Proof. The pre-softmax attention score matrix is computed as $\mathbf{S} = \mathbf{H}\mathbf{W}_q\mathbf{W}_k^\top\mathbf{H}^\top$. By substituting the SVD of $\mathbf{H}$ and defining a projection core matrix $\mathbf{M} = \mathbf{V}^\top\mathbf{W}_q\mathbf{W}_k^\top\mathbf{V} \in \mathbb{R}^{r \times r}$, we can express $\mathbf{S}$ as:

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}(\mathbf{V}^\top\mathbf{W}_q\mathbf{W}_k^\top\mathbf{V})\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{M}\mathbf{\Sigma}\mathbf{U}^\top. \quad (10)$$

The (i, j) -th entry of $\mathbf{S}$ can be explicitly expanded and decomposed into:

$$\begin{aligned} S_{ij} &= \sum_{a=1}^r \sum_{b=1}^r u_{ia} \sigma_a M_{ab} \sigma_b u_{jb} \\ &= \underbrace{\sigma_1^2 M_{11} u_{i1} u_{j1}}_{\text{Dominant Component}} + \underbrace{\sum_{(a,b) \neq (1,1)} \sigma_a \sigma_b M_{ab} u_{ia} u_{jb}}_{\text{Minor Component } R_{ij}}. \end{aligned} \quad (11)$$

Since $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices ($|u_{ij}| \leq 1$), and the learned weight matrices $\mathbf{W}_q, \mathbf{W}_k$ inherently have bounded norms due to standard regularization, their product $\mathbf{M} = \mathbf{V}^\top\mathbf{W}_q\mathbf{W}_k^\top\mathbf{V}$ is bounded at a relatively small scale. Consequently, the magnitude of each term is strictly governed by the singular values $\sigma_a \sigma_b$. Under the embedding collapse condition ($\sigma_k \leq \epsilon\sigma_1$ for $k \geq 2$), the residual term is bounded by $|R_{ij}| \leq \mathcal{O}(\epsilon\sigma_1^2)$. Therefore, the pre-softmax

score is heavily dominated by the rank-1 principal component. Worse still, when passed through the exponential function $\mathbf{A} = \text{softmax}(\mathbf{S})$, this dominance is exponentially amplified. The attention matrix $\mathbf{A}$ loses sensitivity to minor feature interactions and rapidly collapses into a low-rank matrix governed primarily by $\mathbf{u}_1$. $\square$

Proposition III.2 (Backward Propagation: Gradient Concentration on Dominant Components Amplifies Collapse). *Given the smoothed and degenerate attention matrix $\mathbf{A}$ derived from Proposition III.1, the gradients during backpropagation are heavily concentrated on the dominant spectral component, while minor components receive negligible updates, rendering the model incapable of recovering lost feature diversity.*

Proof. Consider the backward propagation through the value aggregation phase $\mathbf{O} = \mathbf{A}\mathbf{X}$ (where $\mathbf{X} = \mathbf{H}\mathbf{W}_v$). (While gradients also backpropagate through the query and key projections, those updates inherently involve right-multiplications with the highly rank-deficient input $\mathbf{H}$, meaning their gradients are also strictly confined to the dominant spectral subspace [38]). Focusing on the gradient path through the value aggregation, by the chain rule, the gradient with respect to the input representations $\mathbf{X}$ is:

$$\nabla_{\mathbf{X}} \mathcal{L} = \mathbf{A}^\top \nabla_{\mathbf{O}} \mathcal{L}. \quad (12)$$

From Proposition III.1, the pre-softmax score matrix can be expressed as $\mathbf{S} = \mathbf{S}^1 + \mathbf{R}$, where $\mathbf{S}^1 = \sigma_1^2 M_{11} \mathbf{u}_1 \mathbf{u}_1^\top$ is the rank-1 dominant component and the residual satisfies $\|\mathbf{R}\|_\infty = \mathcal{O}(\epsilon\sigma_1^2)$. Let $\mathbf{A}^1 = \text{softmax}(\mathbf{S}^1)$. Due to the Lipschitz continuity of the softmax function [39], the actual attention matrix satisfies $\mathbf{A} = \mathbf{A}^1 + \mathbf{A}^E$, where the error matrix $\mathbf{A}^E$ is bounded by $\|\mathbf{A}^E\| \leq \mathcal{O}(\epsilon\sigma_1^2)$.

To see how this affects the gradient flow, we project the upstream gradient $\nabla_{\mathbf{O}} \mathcal{L}$ onto the orthogonal basis $\mathbf{U}$:

$$\nabla_{\mathbf{O}} \mathcal{L} = \mathbf{u}_1 \mathbf{g}_1^\top + \sum_{k=2}^r \mathbf{u}_k \mathbf{g}_k^\top, \quad (13)$$

where $\mathbf{g}_k$ represents the gradient vector corresponding to the k -th spectral component. The backward update can then be expanded as:

$$\nabla_{\mathbf{X}} \mathcal{L} = (\mathbf{A}^1 + \mathbf{A}^E)^\top \left(\mathbf{u}_1 \mathbf{g}_1^\top + \sum_{k=2}^r \mathbf{u}_k \mathbf{g}_k^\top \right). \quad (14)$$

As proven by the previous work [38], the self-attention matrix acts intrinsically as a low-pass filter. Because $\mathbf{A}^1$ is generated by the latent structure of $\mathbf{u}_1$, its rows exhibit minimal variance and lack high-frequency complexity. Consequently, when operating on the orthogonal basis $\mathbf{U}$, the transformation $\mathbf{A}^{1\top}$ strongly preserves the primary structural component $\mathbf{u}_1$, but severely attenuates the high-frequency minor components $\mathbf{u}_k$ ($k \geq 2$).

Let $\xi_k = \|\mathbf{A}^{1\top} \mathbf{u}_k\| / \|\mathbf{u}_k\|$ denote the attenuation factor, which quantifies the fraction of the k -th spectral component's magnitude preserved during backpropagation. For minor spectral components, the low-pass filtering property dictates that

$\xi_k \ll \xi_1$. Thus, the gradient update projected onto the k -th minor component ($k \geq 2$) is bounded by:

$$\|\nabla_{\mathbf{X}} \mathcal{L} \cdot \mathbf{u}_k\| \leq \xi_k \|\mathbf{g}_k\| + \mathcal{O}(\epsilon \sigma_1^2). \quad (15)$$

Since the attenuation factor ξ_k is exceedingly small and the residual perturbation is bounded by $\mathcal{O}(\epsilon \sigma_1^2)$, the gradient is overwhelmingly concentrated on the dominant component $\mathbf{u}_1$. The minor components receive highly attenuated updates, which exacerbates the embedding collapse and prevents the model from escaping this degenerate state. $\square$

The Vicious Cycle of Collapse in Recommendation.

Propositions III.1 and III.2 collectively reveal a closed-loop vicious cycle: initially collapsed embeddings in recommendation generate smoothed, low-rank attention maps (Forward Propagation). These degenerate attention maps, acting as low-pass filters, strictly restrict the gradient updates to the dominant spectral directions (Backward Propagation). This pushes the embeddings to become even more collapsed in the next iteration. This fundamental mechanism mathematically explains why standard self-attention suffers from attention collapse much faster and more severely in recommendation scenarios compared to LLMs.

IV. METHODOLOGY

To address the embedding collapse and attention collapse in recommendation, we propose **SpecFormer**, a Spectral-Aware Transformer tailored for recommendation. As illustrated in Figure 4, each SpecFormer layer replaces the standard self-attention module with three tightly coupled spectral components while keeping the remaining Transformer components unchanged. We describe each component in detail below.

A. Learnable Spectral Softening

The root cause of embedding collapse is the skewed singular value distribution of the feature embedding matrix, where a few dominant singular values capture nearly all the representational information. We directly address this by applying a learnable spectral softening to the input at each layer before attention computation.

Formally, let $\mathbf{H}^{(l)} \in \mathbb{R}^{L \times D}$ denote the input to the l -th Transformer layer, with $\mathbf{H}^{(0)} = \mathbf{X}$. We compute the SVD as:

$$\mathbf{H}^{(l)} = \mathbf{U}^{(l)} \mathbf{\Sigma}^{(l)} \mathbf{V}^{(l)\top}, \quad (16)$$

where $\mathbf{U}^{(l)} \in \mathbb{R}^{L \times r}$, $\mathbf{V}^{(l)} \in \mathbb{R}^{D \times r}$, $r = \min(L, D)$, and $\mathbf{\Sigma}^{(l)} = \text{diag}(\sigma_1^{(l)}, \dots, \sigma_r^{(l)})$ with $\sigma_1^{(l)} \geq \dots \geq \sigma_r^{(l)} \geq 0$.

We then apply a *power-law softening* to the singular value spectrum:

$$\mathbf{\Sigma}^{(l)*} = \text{diag}(\sigma_1^{(l)\tau}, \sigma_2^{(l)\tau}, \dots, \sigma_r^{(l)\tau}), \quad (17)$$

where $\tau = \text{Sigmoid}(a) \in (0, 1)$ and a is a learnable scalar initialized to 0 (*i.e.*, τ starts at 0.5). Since $\tau < 1$, the power-law transformation compresses large singular values more aggressively than small ones, effectively flattening the spectrum and redistributing representational energy across all

dimensions. The softened feature matrix is then reconstructed as:

$$\mathbf{H}^{(l)*} = \mathbf{U}^{(l)} \mathbf{\Sigma}^{(l)*} \mathbf{V}^{(l)\top}. \quad (18)$$

The power-law softening provides a continuous, data-adaptive control over the degree of spectral flattening via the learnable parameter τ .

B. Spectrum-softened Attention

Standard self-attention computes feature interactions directly in the original feature space, which causes attention to be dominated by the collapsed principal component. We instead propose to compute attention weights in the *spectrum-softened domain* with the softened embeddings, so that all spectral components contribute meaningfully to feature interactions.

Specifically, we define the Query and Key matrices using the softened embedding $\mathbf{H}^{(l)*}$, while keeping the Value matrix in the original feature space:

$$\mathbf{Q}^{(l)} = \mathbf{H}^{(l)*} \mathbf{W}_q^{(l)}, \quad \mathbf{K}^{(l)} = \mathbf{H}^{(l)*} \mathbf{W}_k^{(l)}, \quad \mathbf{V}_{\text{val}}^{(l)} = \mathbf{H}^{(l)}, \quad (19)$$

where $\mathbf{W}_q^{(l)}, \mathbf{W}_k^{(l)} \in \mathbb{R}^{D \times D}$. Using the original $\mathbf{H}^{(l)}$ (without softening or projection) as the Value matrix is motivated empirically: the Value matrix aggregates feature content, and preserving the original representation avoids information loss in the output. We also compare it with other Value matrix designs and find it performs best (*cf.* Section V-C).

The pre-softmax attention score matrix is:

$$\mathbf{S}^{(l)} = \mathbf{Q}^{(l)} \mathbf{K}^{(l)\top} = \mathbf{H}^{(l)*} \mathbf{W}_q^{(l)} \mathbf{W}_k^{(l)\top} \mathbf{H}^{(l)*\top}. \quad (20)$$

To reveal the spectral nature of this computation, we substitute Eq. (18) into Eq. (20) and drop the layer superscript for clarity:

$$\mathbf{S} = \mathbf{U} \underbrace{(\mathbf{\Sigma}^* \mathbf{V}^\top \mathbf{W}_q \mathbf{W}_k^\top \mathbf{V} \mathbf{\Sigma}^*)}_{\text{spectrum-softened attention}} \mathbf{U}^\top. \quad (21)$$

This design naturally prevents a single spectral component from dominating the attention computation, as the softened spectrum $\mathbf{\Sigma}^*$ ensures a more balanced contribution from all components.

C. Spectral Residual Position Encoding

In addition to the spectrum-softened attention, we further introduce a *learnable spectral residual position encoding* to provide an explicit spectral inductive bias for feature interactions and prevent collapse from compounding across layers.

Spectral Residual Term. We define a spectral position encoding bias $\mathbf{P}_{\text{bias}}^{(l)} \in \mathbb{R}^{L \times L}$ as:

$$\mathbf{P}_{\text{bias}}^{(l)} = \mathbf{U}^{(l)} \mathbf{P}^{(l)} \mathbf{U}^{(l)\top}, \quad (22)$$

where $\mathbf{P}^{(l)} \in \mathbb{R}^{r \times r}$ is a learnable spectral interaction weight matrix. To parameterize $\mathbf{P}^{(l)}$ in a structured and data-adaptive

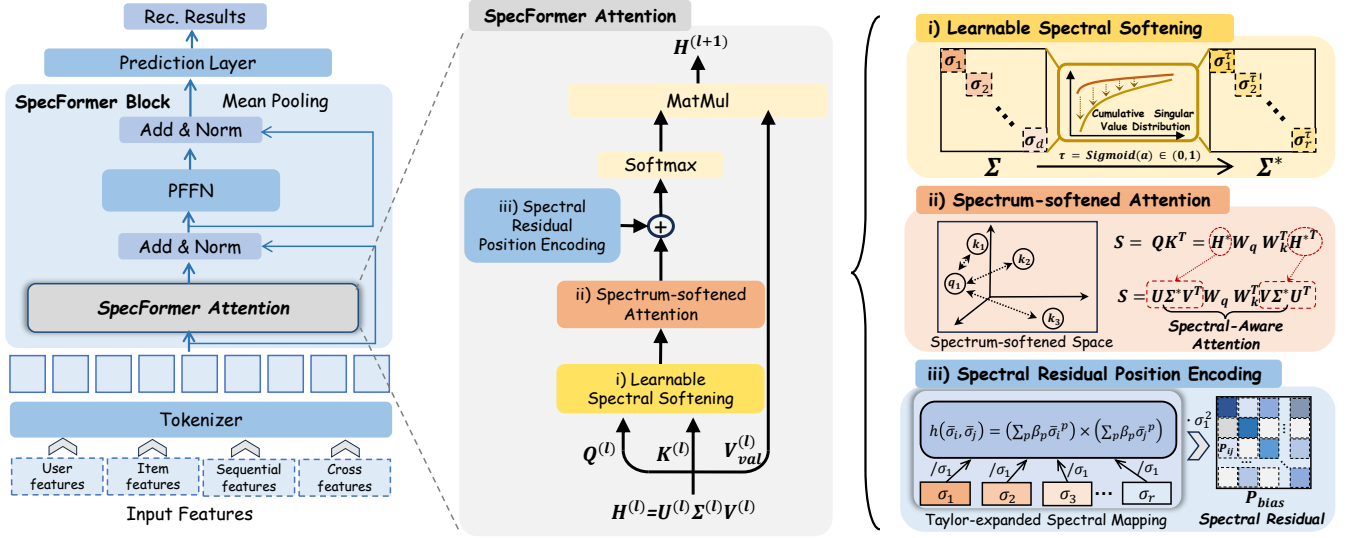


Fig. 4: The architecture of SpecFormer. Each SpecFormer attention layer consists of three components: (1) Learnable Spectral Softening applied to the input feature matrix; (2) Spectrum-softened Attention operating in the spectral-aware feature interaction space; and (3) Spectral Residual Position Encoding providing a learnable Taylor-expanded spectral bias term.

manner, we define its (i, j) -th entry via a *Taylor-expanded pairwise spectral mapping*:

$$\bar{\sigma}_i = \sigma_i / \sigma_1, \quad (23)$$

$$h(\bar{\sigma}_i, \bar{\sigma}_j) = \left(\sum_{p=0}^2 \beta_p \bar{\sigma}_i^p \right) \times \left(\sum_{p=0}^2 \beta_p \bar{\sigma}_j^p \right), \quad (24)$$

$$P_{ij}^{(l)} = \sigma_1^2 h(\bar{\sigma}_i, \bar{\sigma}_j), \quad (25)$$

where $\bar{\sigma}_i$ is the normalized singular value, $h(\bar{\sigma}_i, \bar{\sigma}_j)$ is a second-order polynomial that models the pairwise spectral interaction between the i -th and j -th feature components, and $\beta = [\beta_0, \beta_1, \beta_2]^\top$ are shared learnable parameters.

Crucially, this specific parameterization provides a vital structural inductive bias to compensate for the original spectral weights. Because the spectral softening (Section IV-A) effectively flattens the representation space to revitalize tail features, it might inadvertently penalize the highly informative dominant components. By explicitly anchoring the interactions on the maximum singular value σ_1^2 and the normalized spectrum, this bias term acts as a targeted structural residual. It preserves the heavy-weighted principal spectral signals, preventing the loss of main recommendation patterns in the uniformized space.

Spatial Residual Term. To further stabilize training and preserve the feature interaction signal from the original space, we add a second residual term computed from the un-softened embeddings:

$$\mathbf{S}_{\text{bias}}^{(l)} = \left(\mathbf{H}^{(l)} \tilde{\mathbf{W}}_q^{(l)} \right) \left(\mathbf{H}^{(l)} \tilde{\mathbf{W}}_k^{(l)} \right)^\top, \quad (26)$$

where $\tilde{\mathbf{W}}_q^{(l)}$ and $\tilde{\mathbf{W}}_k^{(l)}$ are independent linear projections dedicated to the residual connection, ensuring the gradient flows of the original and softened spaces are independent.

Final SpecFormer Attention. Combining the spectral attention score (Eq. 20), the spectral bias (Eq. 22), and the spatial residual (Eq. 26), the complete SpecFormer attention is:

$$\text{SpecAtt}(\mathbf{H}^{(l)}) = \phi \left(\mathbf{S}^{(l)} + \alpha \mathbf{P}_{\text{bias}}^{(l)} + \gamma \mathbf{S}_{\text{bias}}^{(l)} \right) \mathbf{V}_{\text{val}}^{(l)}, \quad (27)$$

where $\phi(\cdot) = \text{softmax}(\cdot)$ is the activation function, and $\alpha, \gamma \in \mathbb{R}$ are hyperparameters that control the strength of each residual term.

The remaining components follow the standard Transformer design. The full update rule for the l -th SpecFormer layer is:

$$\mathbf{Z}^{(l)} = \text{LN} \left(\text{SpecAtt}(\mathbf{H}^{(l)}) + \mathbf{H}^{(l)} \right), \quad (28)$$

$$\mathbf{H}^{(l+1)} = \text{LN} \left(\mathbf{Z}^{(l)} + \text{PFFN}(\mathbf{Z}^{(l)}) \right), \quad (29)$$

where $\text{LN}(\cdot)$ is Layer Normalization [29] and $\text{PFFN}(\cdot)$ is the pre-token FFN introduced in Section II. The final representation is mean-pooled over the feature dimension and passed to a prediction MLP.

V. EXPERIMENTS

In this section, we conduct extensive experiments to evaluate the proposed SpecFormer. Specifically, we aim to answer the following research questions:

- **RQ1:** How does SpecFormer perform compared to existing state-of-the-art traditional, Transformer-based, and spectral-based recommendation models?
- **RQ2:** How do the key components of SpecFormer contribute to the overall performance?
- **RQ3:** Does SpecFormer mitigate the collapse bottleneck and achieve scaling capability on stacking Transformer layers?
- **RQ4:** What are the impacts of different hyperparameters in SpecFormer?

TABLE II: Statistics of the datasets used in experiments.

Dataset	#Samples	#Fields	#Features	#Positive
Industrial	1.2B	391	>1B	2.7%
Criteo	45M	39	1.33M	26.0%
Avazu	40M	23	1.54M	17.0%

- **RQ5:** How does SpecFormer perform in a real-world, large-scale online industrial system in terms of system efficiency and business metrics?

A. Experimental Setup

1) *Datasets:* Following previous work [11], [14], we evaluate SpecFormer on the following datasets: one industrial dataset and two well-known public benchmarks — Criteo¹, and Avazu². Table II summarizes the statistics of each dataset.

- **Industrial Dataset** is collected from the online display advertising system of a leading international e-commerce platform. It contains approximately 1.2 billion exposure samples with user behavior sequences up to a length of 256, containing 38 feature fields.
- **Criteo** is a well-known public benchmark for CTR prediction. This dataset contains one week of real-world advertisement click data. It features a mix of 13 continuous (numerical) features and 26 categorical features.
- **Avazu** is a widely-used benchmark consisting of 10 days of chronologically ordered ad click logs. Each sample contains 23 feature fields, providing a robust environment for testing on rich time-ordered categorical data.

2) *Baselines:* We compare SpecFormer against three categories of baselines:

- **Traditional Feature Interaction Models:** **DeepFM** (arXiv’17) [23] integrates factorization machines and deep neural networks to jointly model both low-order and high-order feature interactions. **AutoInt** (CIKM’19) [10] utilizes a multi-head self-attention to explicitly map features into a projected space for automated high-order interaction learning. **MaskNet** (arXiv’21) [40] introduces an instance-guided mask module to dynamically highlight salient features and capture complex non-linear feature interactions. **FiBiNet++** (CIKM’23) [24] enhances the FiBiNet [20] by refining the bilinear interaction process and attention mechanisms for more efficient representation learning. **GDCN** (CIKM’23) [41] proposes a gated cross network that captures explicit high-order feature interactions while utilizing gating mechanisms to dynamically filter out noise.
- **Transformer-Based Feature Interaction Models:** **HiFormer** (arXiv’23) [28] designs a hierarchical Transformer architecture to effectively capture both fine-grained intra-feature and coarse-grained inter-feature relationships. **FAT** (arXiv’25) [16] scales up feature interaction learning by adapting a comprehensive Transformer architecture to model complex, long-range cross-feature dependencies.

RankMixer (CIKM’25) [11] integrates a specialized token-mixing attention mechanism tailored for ranking tasks to efficiently aggregate and transform feature interactions. **One-Trans** (WWW’26) [6] proposes a Transformer framework to seamlessly integrate various feature interaction paradigms into a single unified attention mechanism.

- **Spectral-Based Recommendation Models:** **WhitenRec** (ICDE’24) [17] applies a spectral-based whitening transformation to alleviate dimensional collapse. **FEDIN** (arXiv’26) [18] leverages Fourier Transform to optimize feature representations in the frequency domain and model global interaction patterns.

3) *Evaluation Metrics:* Following previous work [6], [11], We employ the widely used **AUC** (Area Under the ROC Curve) and **Logloss** (Cross-Entropy Loss) as standard evaluation metrics. For the industrial dataset, we additionally report **GAUC** (Group AUC), which measures user-specific ranking quality and is highly correlated with online performance. Note that a **0.001** improvement in AUC/GAUC is considered significant and can lead to massive revenue growth in industrial applications [10], [42].

4) *Implementation Details:* We implement all models in PyTorch and conduct experiments on servers equipped with NVIDIA L20 GPUs. For the industrial dataset, the token embedding dimension is set to 304, optimized via Adagrad [43] with a batch size of 1024. For public datasets (Criteo, and Avazu), the embedding dimensions are configured as 312, and 176, respectively, using a batch size of 512. The optimal learning rate for each model is determined through a grid search within $\{0.005, 0.01, 0.05, 0.1\}$ for the industrial dataset and a broader range of $\{0.001, 0.002, 0.005, 0.01, 0.05, 0.1\}$ for public datasets. Our code is available at <https://github.com/istarryn/SpecFormer>.

5) *Training Strategy and Warm-up Protocol:* To ensure numerical stability and model convergence, we adopt a two-stage training protocol for SpecFormer, whereas the baseline models follow a standard single-stage procedure: (a) *Baseline Configuration:* Following the common practice in CTR prediction, all baseline models are initialized using a standard Gaussian distribution and trained on the full training set (100%) to optimize the target CTR loss. (b) *SpecFormer Configuration:* We employ a two-stage training scheme. In the first stage (Embedding Warm-up), 5% of the training data is utilized to initialize the embeddings by optimizing the *Spatial Residual Term*. In the second stage (Main CTR Training), the model is trained on the remaining 95% of the data using the complete architecture, including the Spectral Softening module.

The warm-up phase is essential for stabilizing the *Spectral Softening* mechanism. Randomly initialized embeddings often exhibit ill-conditioned singular value distributions dominated by noise, making direct spectral transformations prone to numerical instability. By pre-training with the *Spatial Residual Term*, we project embeddings onto a structured feature manifold. This establishes a stable spectral foundation, enabling

¹<https://www.kaggle.com/c/criteo-display-ad-challenge/data>

²<https://www.kaggle.com/c/avazu-ctr-prediction/data>

robust frequency-domain refinement during the subsequent CTR training stage.

B. Performance Comparison (RQ1)

Table III shows the performance comparison of the proposed SpecFormer against the baseline methods. We observe that:

1) Overall performance comparisons. SpecFormer consistently and significantly outperforms all baseline models across the evaluated datasets in terms of AUC, GAUC, and Logloss. These impressive results validate the effectiveness and generalizability of our proposed spectral-aware architecture. By fundamentally addressing the spectral collapse bottleneck, SpecFormer proves to be a highly effective and robust feature interaction backbone for real-world recommender systems.

2) Compared with Traditional Feature Interaction Models. SpecFormer significantly outperforms traditional recommendation architectures. While these traditional models capture low- and high-order feature interactions through well-designed structures, they inherently lack the context-aware modeling capabilities to model complex dependencies. Furthermore, although models like AutoInt attempt to employ early self-attention, they operate directly on raw feature embeddings, making them highly susceptible to parameter fragmentation [11]. In contrast, SpecFormer utilizes a modernized tokenization paradigm and operates within the spectral attention, unlocking a much larger capacity for expressive feature interactions.

3) Compared with Transformer-based baselines. While modern Transformer-based models generally outperform traditional architectures, they still hit a noticeable performance ceiling. Notably, models adopting the standard self-attention mechanism (e.g., OneTrans) sometimes even underperform well-designed but simpler token-mixing methods (e.g., RankMixer). This empirical evidence strongly corroborates our theoretical analysis in Section III: when faced with heterogeneous recommendation data, standard attention mechanisms cause the attention maps to be dominated by a few principal singular values. In contrast, SpecFormer surpasses these strong state-of-the-art baselines by modeling feature interactions within a dynamically softened spectral space. The spectral-aware design effectively revitalizes tail features, forcing the model to capture a more balanced and diverse set of spectral signals, thereby breaking the bottleneck caused by the standard attention mechanism.

4) Compared with Spectral-based baselines. SpecFormer exhibits substantial superiority over the spectral-based baselines. Specifically, WhitenRec adjusts the spectrum using static SVD and spectral constraints, but normal spectral transformations are insufficient for complex feature interactions. Furthermore, FEDIN performs poorly across the datasets, confirming that simple frequency-domain transformations struggle to capture non-linear attention-based feature interactions. Conversely, SpecFormer intrinsically integrates spectral-aware feature interaction and mitigates the core spectral collapse in attention mechanism, achieving better performance.

C. Ablation Study (RQ2)

We perform the following ablation study to investigate the effects of each component in SpecFormer:

- **1) Learnable Spectral Softening Module.** (a) w/o Spectral Softening ($\Sigma^* = \Sigma$): We remove the learnable power-law softening and directly use the original singular values to reconstruct the space.
- **2) Value Matrix Design.** (b) Value Matrix w/ Projection ($V_{\text{val}} = HW_v$): We apply a standard linear projection to the Value matrix. (c) Value Matrix w/ Softening ($V_{\text{val}} = H^*$): We use the spectrum-softened embeddings as the Value matrix instead of the original design.
- **3) Spectral and Spatial Residuals.** (d) w/o Spectral Position Encoding ($\alpha = 0$): We remove the Taylor expansion-based spectral residual position encoding bias. (e) Spectral Position Encoding $\rightarrow$ Identity Matrix ($P = I$): We replace the learnable Taylor expansion-based spectral mapping matrix with an identity matrix. (f) w/o Spatial Residual ($\gamma = 0$): We remove the spatial residual term S_{bias} .
- **4) Spectrum-softened Attention Mechanism.** (g) w/o Spectrum-softened Attention S (Only Residuals): We remove the core dynamic spectrum-softened attention matrix S, leaving only the structural residual biases.

As shown in Table IV, we can observe that:

1) Removing the *Learnable Spectral Softening* leads to a catastrophic performance decline, particularly on the most important AUC metric. This directly validates our core motivation (cf. Section III): without dynamically flattening the singular value distribution, the attention mechanism immediately collapses into the dominant principal components, losing its ability to capture long-tail feature interactions.

2) Regarding the *Value Matrix Design*, both (b) and (c) yield severely sub-optimal performance. Applying a standard projection W_v to the original collapsed input H is unnecessary and leads to sub-optimal performance. Similarly, using the softened representation H^* as the Value matrix artificially distorts the semantics of the embedding and might cause information loss. This exploration empirically justifies our design in Section IV-B: utilizing the original H as the Value matrix effectively preserves the raw information, while the softened Query and Key matrices model spectral-aware feature interactions.

3) The ablation on *Residual Position Encodings* demonstrates the necessity of preserving original spectral and spatial signals. Removing the Spectral Position Encoding causes a noticeable drop, proving that while spectral softening revitalizes tail features, it inevitably penalizes principal spectral patterns. The Spectral Position Encoding is crucial as a structural residual to compensate for these essential signals. Furthermore, replacing it with an Identity Matrix results in one of the most severe performance drops. A completely uniform distribution fails to anchor the interactions to the maximum spectrum (σ_1^2), making it hard for the model to differentiate the varying importance of distinct spectral components. Finally, removing the Spatial Residual slightly degrades performance,

TABLE III: Performance comparisons of SpecFormer with different representative baselines on three real-world datasets. A higher AUC/GAUC and a lower Logloss indicate better performance. The best and the second performances are **bolded** and underlined, respectively.

Category	Model	Industrial			Criteo		Avazu	
		AUC $\uparrow$	GAUC $\uparrow$	Logloss $\downarrow$	AUC $\uparrow$	Logloss $\downarrow$	AUC $\uparrow$	Logloss $\downarrow$
Traditional	Wide&Deep	0.7373	0.6287	0.1144	0.6589	0.5372	0.6364	0.4335
	DeepFM	0.7378	0.6282	0.1143	0.5988	0.5421	0.7404	0.3965
	GDCN	0.7491	0.6426	0.1133	0.6727	0.5268	0.7370	0.3989
	AutoInt	0.7483	0.6405	0.1136	0.7022	0.5153	0.7439	0.3952
	FiBiNet++	0.7508	0.6437	0.1133	0.7485	0.4868	0.7410	0.3965
Transformer-based	HiFormer	0.7497	0.6417	0.1134	<u>0.7570</u>	<u>0.4811</u>	0.7248	0.4842
	FAT	0.7559	0.6490	<u>0.1126</u>	0.7551	0.4828	0.7440	0.3955
	RankMixer	<u>0.7587</u>	<u>0.6526</u>	0.1127	0.7484	0.4859	0.7450	0.3951
	OneTrans	0.7555	0.6497	0.1127	0.7558	0.4823	<u>0.7461</u>	0.3943
Spectral-based	WhitenRec	0.7481	0.6425	0.1139	0.7472	0.4888	0.7484	<u>0.3932</u>
	FEDIN	0.6945	0.5675	0.1189	0.6172	0.9568	0.6737	0.4256
SpecFormer		0.7611	0.6537	0.1119	0.7587	0.4791	0.7574	0.3902

highlighting its role in stabilizing training from the original spatial perspective.

4) Removing the *Spectrum-softened Attention S* leads to a significant decline across all datasets. This indicates that while the residual biases provide excellent structural priors, they cannot replace the dynamic and data-dependent feature interaction capabilities of the attention mechanism. SpecFormer achieves state-of-the-art performance exclusively through the synergistic combination of the core spectrum-softened attention mechanism and the structured residuals.

D. Scaling Study (RQ3)

To evaluate the scalability of SpecFormer against other Transformer-based models on the depth of Transformer layers, we conducted a comparative study against the representative models RankMixer and OneTrans by scaling the model depth from 3 to 5, 7, and 9 layers on the Industrial dataset. We analyzed the performance variations through two key dimensions: the incremental AUC gains (Δ AUC) and the layer-wise fractional effective rank of the attention matrices. As shown in Figure 5, we observe that:

1) As the number of Transformer layers increases, the AUC of SpecFormer and RankMixer consistently grows, whereas OneTrans exhibits a severe rise-then-fall performance degradation. This corroborates our theoretical analysis in Section III: OneTrans, relying on standard self-attention, inevitably falls into the vicious cycle of attention collapse. Conversely, SpecFormer fundamentally breaks this bottleneck, demonstrating the capacity to continuously scale up on model depth for substantial recommendation performance gains.

2) While both SpecFormer and RankMixer show the scaling capacity, SpecFormer exhibits a steeper AUC growth curve and consistently outperforms RankMixer across 3, 5, 7, and 9 layers. This observation proves that while carefully hand-crafted

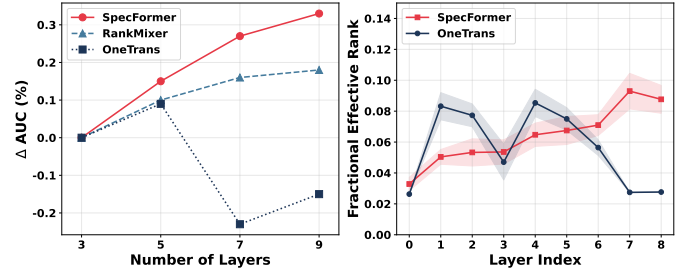


Fig. 5: Comparison of AUC gains and attention effective ranks across Transformer layer depth on RankMixer, OneTrans and SpecFormer.

token-mixing strategies in RankMixer can avoid the attention collapse by simplifying interactions, they inherently lack the expressive power of complex feature modeling. SpecFormer, equipped with the spectral-aware attention mechanism, not only mitigates the collapse but also unleashes the potential of Transformers in recommendation, modeling fine-grained and high-order feature interactions far more effectively than static token-mixing strategies.

3) As model depth increases, the fractional effective rank of SpecFormer’s attention matrix exhibits a steady upward trend. In sharp contrast, the attention effective rank of OneTrans experiences violent fluctuations in early layers and irreversible collapse at deeper layers. This provides the empirical proof for our method: through the spectral-aware attention design, SpecFormer effectively breaks the loop of original attention collapse. It ensures that even at deep depths, the attention map retains a high effective rank, successfully capturing feature interactions across diverse spectral components.

TABLE IV: Ablation study of SpecFormer.

Method	Industrial			Criteo		Avazu	
	AUC $\uparrow$	GAUC $\uparrow$	Logloss $\downarrow$	AUC $\uparrow$	Logloss $\downarrow$	AUC $\uparrow$	Logloss $\downarrow$
SpecFormer	0.7611	0.6537	0.1119	0.7587	0.4791	0.7574	0.3902
(a) w/o Spectral Softening ($\Sigma^* = \Sigma$)	0.7576	0.6541	0.1100	0.6555	0.5332	0.7525	0.3914
(b) Value Matrix w/ Projection ($\mathbf{V}_{\text{val}} = \mathbf{H}\mathbf{W}_v$)	0.7555	0.6526	0.1103	0.6593	0.5351	0.5645	0.4476
(c) Value Matrix w/ Softening ($\mathbf{V}_{\text{val}} = \mathbf{H}^*$)	0.7547	0.6518	0.1100	0.6472	0.5346	0.5643	0.4479
(d) w/o Spectral Position Encoding ($\alpha = 0$)	0.7588	0.6555	0.1099	0.6598	0.5299	0.6700	0.4169
(e) Spectral Position Encoding $\rightarrow$ Identity Matrix ($\mathbf{P} = \mathbf{I}$)	0.7573	0.6548	0.1101	0.7536	0.4832	0.5667	0.4469
(f) w/o Spatial Residual ($\gamma = 0$)	0.7551	0.6527	0.1106	0.7555	0.4812	0.6733	0.4203
(g) w/o Spectrum-softened Attention $\mathbf{S}$ (Only Residuals)	0.7582	0.6551	0.1100	0.7569	0.4812	0.7382	0.3980

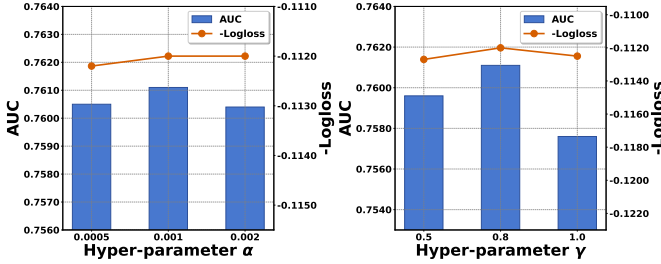


Fig. 6: Hyperparameter sensitivity analysis of SpecFormer.

E. Hyperparameter Sensitivity (RQ4)

Figure 6 showcases the AUC and Logloss performance of the proposed SpecFormer on the Industrial dataset under different spectral and spatial residual position encoding weights (*i.e.*, α and γ). We observe that: 1) As α increases from 0.0001 to 0.001 and 0.01, the recommendation performance initially improves and subsequently declines, generally peaking around 0.001. Similarly, as γ scales across $\{0.5, 0.8, 1.0\}$, the performance exhibits a clear rise-then-fall pattern, achieving optimal results at approximately 0.8. This indicates that an appropriate weight of spectral and spatial bias is essential to compensate for the original spectral and attention signals. However, an excessively large α or γ will overwhelm the dynamically learned spectrum-softened attention and degrade recommendation performance. 2) Notably, the optimal value of α (0.001) is orders of magnitude smaller than that of γ (0.8). The reasons are: (a) The spatial residual $\mathbf{S}_{\text{bias}}^{(l)}$ is calculated via standard dot-product attention, placing its numerical scale close to the primary spectrum-softened attention matrix $\mathbf{S}^{(l)}$. Consequently, γ naturally operates at a scale around 0.8. (b) The spectral residual term $\mathbf{P}_{\text{bias}}^{(l)}$ is explicitly parameterized by the polynomial expansion of singular values and anchored by σ_1^2 . Its numerical values are intrinsically much larger than the standard pre-softmax attention scores. Therefore, a strictly small α is required to regularize the magnitude, preventing gradient explosion and ensuring stable training.

F. Online Experiment (RQ5)

Model Scaling and Multi-epoch Training Strategy. SpecFormer exhibits superior scalability through its deep ar-

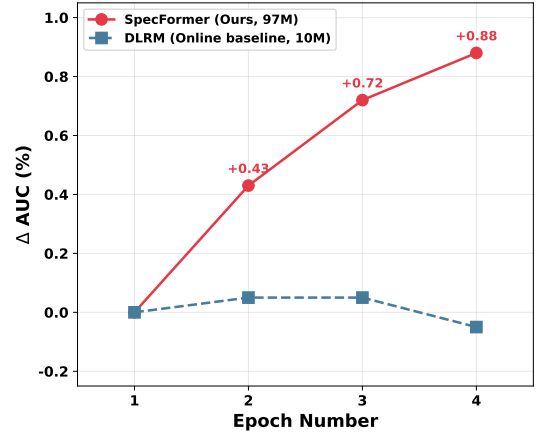


Fig. 7: The scaling capacity of SpecFormer against online DLRM baseline under multi-epoch training.

chitecture, making it highly effective for large-scale parameter expansion. To fully exploit its increased model capacity, we implement a multi-epoch training strategy with a progressive parameter inheritance mechanism. Unlike traditional models like DLRM, which often suffer from rapid performance saturation, SpecFormer continues to yield significant AUC gains as the number of training epochs increases. This is achieved by utilizing warmed-up embeddings at the start of each epoch and inheriting optimized parameters from the previous one, thereby effectively amplifying the benefits of model scaling as shown in Figure 7. To balance training costs and efficiency, we deployed the 3-epoch model online, achieving a significant 0.92% AUC gain over the production DLRM baseline.

Efficiency Optimization. To address the computational complexity of SVD, we employ an in-batch SVD operator that balances spectral refinement with training throughput. On the system level, we integrate a Distributed Dynamic Embedding (DDE) framework, where embedding tables are sharded across different ranks. This allows for dynamic expansion of the feature space while significantly reducing memory footprint. To mitigate communication bottlenecks caused by high-dimensional features, we utilize coalesced lookup operations, successfully reducing embedding query frequency to

TABLE V: Results of the online A/B experiment on a large-scale e-commerce platform. All performance gains are statistically significant with $p < 0.05$.

Method	CTR	CVR	Order	Δ Latency
DLRM Baseline	-	-	-	-
SpecFormer	+1.34%	+15.97%	+16.72%	+5ms

8% of the baseline. For computational efficiency, we integrate FlashAttention, which reduces per-layer attention overhead by 40%. Collectively, these optimizations, combined with FP16 mixed-precision training, deliver an overall 35% improvement in training efficiency.

Online Performance. We conducted a large-scale A/B test on Alibaba’s e-commerce advertising platform to evaluate the real-world efficacy of SpecFormer. The experiment was carried out over a 10-day period from June 1 to June 10, 2026, involving 10% of the platform’s production traffic. Users were randomly assigned to either the control group (a highly optimized DLRM baseline) or the experimental group (SpecFormer) using an orthogonal hashing protocol. As summarized in Table V, SpecFormer achieved significant improvements across all key business metrics, including CTR (+1.34%), and Order (+16.72%), at the cost of merely a 5ms increase in inference latency, which remains well within the strict online service requirements.

VI. RELATED WORK

A. Transformer-based Recommender Systems

Transformer architectures [1] have achieved remarkable success across various domains [2], [5]. In RS, early adoptions primarily focused on capturing sequential user behaviors [8], [9], [44]. Subsequently, Transformers were introduced into CTR prediction to automatically model complex and high-order feature interactions [10], [45]. To explicitly map features into interaction spaces, early methods like AutoInt [10] apply multi-head self-attention directly to raw feature embeddings. Building upon this, recent work explores various structural enhancements to model interaction dependencies more effectively. For instance, HiFormer [28] utilizes a hierarchical Transformer to handle both fine-grained intra-feature and coarse-grained inter-feature relationships. Furthermore, to optimize interaction efficiency and alleviate parameter fragmentation, modern architectures refine the tokenization and aggregation processes. RankMixer [11] integrates a specialized token-mixing attention tailored for ranking tasks, and OneTrans [6] formulates a unified attention framework to seamlessly incorporate multiple feature interaction paradigms. Besides, LLMs also demonstrate strong potential in RS, either functioning as direct recommenders [46]–[49] or acting as semantic enhancers for recommendation models [50]–[53]. However, due to their prohibitive inference latency, deploying LLMs in large-scale industrial systems remains highly challenging, and they are not the focus of this paper.

Despite their success, the attention mechanism suffers from suboptimal performance due to the heterogeneity and

long-tail distribution characteristics of recommendation data in RS. Existing attempts try to address this with multi-parameters [14], [15] or external spectral enhancement [17], [18], but they either rely heavily on spatial-domain patches or fail to integrate the core attention mechanism, resulting in a limited improvement. In contrast, SpecFormer fundamentally uncovers the limitation of attention mechanism and resolves the performance bottleneck from a spectral perspective.

B. Spectral Collapse in Recommender Systems

Embedding collapse refers to the phenomenon where representations degenerate into a low-dimensional subspace [35], [36], which is a problem extensively studied in Natural Language Generation [54] and Contrastive Learning [55], [56]. However, as recent studies indicate [13], [31], the extreme heterogeneity and long-tail distribution of recommendation data drastically exacerbate this collapse, resulting in a highly skewed embedding space dominated by merely a few principal singular values [32], [57]. Meanwhile, attention collapse occurs when the attention map loses its token-routing discriminability, severely limiting the expressive power of self-attention [38], [58]. Our theoretical analysis (*cf.* Section III) advances this line of research by mathematically formulating the vicious cycle between embedding collapse and attention collapse during both forward and backward propagation, a phenomenon specifically severe in recommendation scenarios. Recent works in recommendation introduce explicit spectral regularizations [17] to prevent dimension collapse, and explore Fourier transforms [18] to capture frequency-domain signals for Transformer-based feature interactions. However, these methods typically treat spectral decomposition either as a post-hoc regularization term or as a static transformation isolated from the core attention module. SpecFormer intrinsically mitigates the embedding and attention collapse with the novel spectral-aware attention, which effectively breaks the vicious cycle of collapse in recommender systems.

VII. CONCLUSION

In this paper, we investigate the bottleneck of Transformer’s attention mechanism in recommender systems. Through empirical and theoretical analysis, we reveal that the intrinsic heterogeneity and long-tail distribution of recommendation data trigger a vicious cycle of spectral collapse in original attention: the collapsed embeddings induce smoothing and low-rank attentions, which in turn act as low-pass filters that starve minor spectral components of gradients during backpropagation. To fundamentally break this bottleneck, we propose SpecFormer, a novel Spectral-Aware Transformer. By introducing a *Learnable Spectral Softening* module, modeling interactions via a *Spectrum-softened Attention*, and the *Spectral Residual Position Encoding*, SpecFormer effectively prevents principal spectral components from dominating the attention map. Extensive offline evaluations and online A/B testing demonstrate that SpecFormer not only achieves state-of-the-art performance but also unlocks the depth-scaling potential of Transformers for recommendation.

REFERENCES

- [1] 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.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.
- [6] Z. Zhang, H. Pei, J. Guo, T. Wang, Y. Feng, H. Sun, S. Liu, and A. Sun, "Onetrans: Unified feature interaction and sequence modeling with one transformer in industrial recommender," in *Proceedings of the ACM Web Conference 2026*, 2026, pp. 8162–8170.
- [7] J. Zhu, Q. Jia, G. Cai, Q. Dai, J. Li, Z. Dong, R. Tang, and R. Zhang, "Final: Factorized interaction layer for ctr prediction," in *Proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval*, 2023, pp. 2006–2010.
- [8] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1441–1450.
- [9] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 197–206.
- [10] W. Song, C. Shi, Z. Xiao, Z. Duan, Y. Xu, M. Zhang, and J. Tang, "Autoint: Automatic feature interaction learning via self-attentive neural networks," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 1161–1170.
- [11] J. Zhu, Z. Fan, X. Zhu, Y. Jiang, H. Wang, X. Han, H. Ding, X. Wang, W. Zhao, Z. Gong *et al.*, "Rankmixer: Scaling up ranking models in industrial recommenders," in *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 2025, pp. 6309–6316.
- [12] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
- [13] S. Lin, C. Gao, J. Chen, S. Zhou, B. Hu, Y. Feng, C. Chen, and C. Wang, "How do recommendation models amplify popularity bias? an analysis from the spectral perspective," in *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, 2025, pp. 659–668.
- [14] X. Guo, J. Pan, X. Wang, B. Chen, J. Jiang, and M. Long, "On the embedding collapse when scaling up recommendation models," *arXiv preprint arXiv:2310.04400*, 2023.
- [15] J. Chen, S. Zhang, B. Hu, C. Zhou, J. Pan, G. Xue, W. Ning, G. Weng, W. Zheng, S. Liu *et al.*, "Rankup: Towards high-rank representations for large scale advertising recommender systems," *arXiv preprint arXiv:2604.17878*, 2026.
- [16] B. Yan, Y. Lei, Z. Zeng, D. Wang, K. Lin, P. Wang, J. Xu, and B. Zheng, "From scaling to structured expressivity: Rethinking transformers for ctr prediction," *arXiv preprint arXiv:2511.12081*, 2025.
- [17] L. Zhang, X. Zhou, Z. Zeng, and Z. Shen, "Are id embeddings necessary? whitening pre-trained text embeddings for effective sequential recommendation," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024, pp. 530–543.
- [18] Z. Dai, J. Wang, J. Pan, D. Liu, L. Xiao, and S.-T. Xia, "Fedin: Frequency-enhanced deep interest network for click-through rate prediction," *arXiv preprint arXiv:2605.01726*, 2026.
- [19] Q. Liu, F. Yu, S. Wu, and L. Wang, "A convolutional click prediction model," in *Proceedings of the 24th ACM international conference on information and knowledge management*, 2015, pp. 1743–1746.
- [20] T. Huang, Z. Zhang, and J. Zhang, "Fibinet: combining feature importance and bilinear feature interaction for click-through rate prediction," in *Proceedings of the 13th ACM conference on recommender systems*, 2019, pp. 169–177.
- [21] J. Pan, J. Xu, A. L. Ruiz, W. Zhao, S. Pan, Y. Sun, and Q. Lu, "Field-weighted factorization machines for click-through rate prediction in display advertising," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1349–1357.
- [22] Y. Sun, J. Pan, A. Zhang, and A. Flores, "Fm2: Field-matrixed factorization machines for recommender systems," in *Proceedings of the web conference 2021*, 2021, pp. 2828–2837.
- [23] 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.
- [24] P. Zhang, Z. Zheng, and J. Zhang, "Fibinet++: Reducing model size by low rank feature interaction layer for ctr prediction," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 4425–4429.
- [25] S. Rendle, "Factorization machines," in *2010 IEEE International conference on data mining*. IEEE, 2010, pp. 995–1000.
- [26] 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.
- [27] 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.
- [28] H. Gui, R. Wang, K. Yin, L. Jin, M. Kula, T. Xu, L. Hong, and E. H. Chi, "Hiformer: Heterogeneous feature interactions learning with transformers for recommender systems," *arXiv preprint arXiv:2311.05884*, 2023.
- [29] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [30] A. Hoecker and V. Kartvelishvili, "Svd approach to data unfolding," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 372, no. 3, pp. 469–481, 1996.
- [31] H. Chen, V. Lai, H. Jin, Z. Jiang, M. Das, and X. Hu, "Towards mitigating dimensional collapse of representations in collaborative filtering," in *Proceedings of the 17th ACM international conference on web search and data mining*, 2024, pp. 106–115.
- [32] Y. Cui, F. Liu, Z. Wang, C. Zhang, J. Wang, C. Wang, and J. Chen, "Spectran: Spectral-aware transformer-based adapter for llm-enhanced sequential recommendation," *arXiv preprint arXiv:2601.21986*, 2026.
- [33] O. Roy and M. Vetterli, "The effective rank: A measure of effective dimensionality," in *2007 15th European signal processing conference*. IEEE, 2007, pp. 606–610.
- [34] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [35] L. Noci, S. Anagnostidis, L. Biggio, A. Orvieto, S. P. Singh, and A. Lucchi, "Signal propagation in transformers: Theoretical perspectives and the role of rank collapse," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 198–27 211, 2022.
- [36] L. Wei, Z. Tan, C. Li, J. Wang, and W. Huang, "Diff-erank: A novel rank-based metric for evaluating large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 39 501–39 521, 2024.
- [37] R. Wang, R. Shivanna, D. Cheng, S. Jain, D. Lin, L. Hong, and E. Chi, "Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems," in *Proceedings of the web conference 2021*, 2021, pp. 1785–1797.
- [38] Y. Dong, J.-B. Cordonnier, and A. Loukas, "Attention is not all you need: Pure attention loses rank doubly exponentially with depth," in *International conference on machine learning*. PMLR, 2021, pp. 2793–2803.
- [39] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *arXiv preprint arXiv:1704.00805*, 2017.
- [40] Z. Wang, Q. She, and J. Zhang, "Masknet: Introducing feature-wise multiplication to ctr ranking models by instance-guided mask," *arXiv preprint arXiv:2102.07619*, 2021.
- [41] F. Wang, H. Gu, D. Li, T. Lu, P. Zhang, and N. Gu, "Towards deeper, lighter and interpretable cross network for ctr prediction," in *Proceedings of the 32nd ACM international conference on information and knowledge management*, 2023, pp. 2523–2533.

- [42] 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.
- [43] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of machine learning research*, vol. 12, no. 7, 2011.
- [44] Z. Yang, X. He, J. Zhang, J. Wu, X. Xin, J. Chen, and X. Wang, "A generic learning framework for sequential recommendation with distribution shifts," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 331–340.
- [45] 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.
- [46] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, and X. He, "Tallrec: An effective and efficient tuning framework to align large language model with recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023, pp. 1007–1014.
- [47] J. Chen, L. Chi, B. Peng, and Z. Yuan, "Hllm: Enhancing sequential recommendations via hierarchical large language models for item and user modeling," *arXiv preprint arXiv:2409.12740*, 2024.
- [48] J. Lin, R. Shan, C. Zhu, K. Du, B. Chen, S. Quan, R. Tang, Y. Yu, and W. Zhang, "Rella: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation," in *Proceedings of the ACM on Web Conference 2024*, 2024, pp. 3497–3508.
- [49] B. Geng, Z. Huan, X. Zhang, Y. He, L. Zhang, F. Yuan, J. Zhou, and L. Mo, "Breaking the length barrier: Llm-enhanced ctr prediction in long textual user behaviors," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2311–2315.
- [50] X. Li, B. Chen, L. Hou, and R. Tang, "Ctrl: Connect collaborative and language model for ctr prediction," *ACM Transactions on Recommender Systems*, vol. 4, no. 2, pp. 1–23, 2025.
- [51] Y. Xi, W. Liu, J. Lin, X. Cai, H. Zhu, J. Zhu, B. Chen, R. Tang, W. Zhang, and Y. Yu, "Towards open-world recommendation with knowledge augmentation from large language models," in *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024, pp. 12–22.
- [52] Z. Sun, Z. Si, X. Zang, K. Zheng, Y. Song, X. Zhang, and J. Xu, "Large language models enhanced collaborative filtering," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 2178–2188.
- [53] Y. Cui, F. Liu, J. Chen, X. Lou, C. Zhang, J. Wang, Y. Sun, X. Yang, and C. Wang, "Field matters: A lightweight llm-enhanced method for ctr prediction," in *Proceedings of the ACM Web Conference 2026*, 2026, pp. 6365–6376.
- [54] J. Gao, D. He, X. Tan, T. Qin, L. Wang, and T.-Y. Liu, "Representation degeneration problem in training natural language generation models," *arXiv preprint arXiv:1907.12009*, 2019.
- [55] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," *arXiv preprint arXiv:2110.09348*, 2021.
- [56] T. Hua, W. Wang, Z. Xue, S. Ren, Y. Wang, and H. Zhao, "On feature decorrelation in self-supervised learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9598–9608.
- [57] G. Hu, A. Zhang, S. Liu, Z. Cai, X. Yang, and X. Wang, "Alphafuse: Learn id embeddings for sequential recommendation in null space of language embeddings," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025, pp. 1614–1623.
- [58] S. Bhojanapalli, C. Yun, A. S. Rawat, S. Reddi, and S. Kumar, "Low-rank bottleneck in multi-head attention models," in *International conference on machine learning*. PMLR, 2020, pp. 864–873.