REACTION: Parameter-Efficient Learning for Recommendation

Song-Li Wu[1], **Zhaocheng Du**[2*], **Qinglin Jia**[2], **Zhenhua Dong**[2]

[1]Tsinghua University
[2]Huawei Noah's Ark Lab
wsl24@mails.tsinghua.edu.cn, zhaochengdu@huawei.com, jiaqinglin2@huawei.com, dongzhenhua@huawei.com

Abstract

While deep learning (DL) has demonstrated significant success in recommender systems, it suffers from high computational complexity and poor scalability. In this work, we demonstrate, from an information-theoretic perspective, the redundancy of existing DL-based recommender models in two aspects: (1) **Feature Redundancy**. We show that many features are highly mutually correlated, noisy, or weakly predictive of user-item interaction labels. (2) **Structural Redundancy**. We further show that a large proportion of parameters in the dense layers contribute minimally to overall performance, indicating significant redundancy within the model architecture. To address these challenges, we propose REACTION (paRameter-Efficient LeArning for recommendaTION), an information-theoretic framework designed to reduce model complexity without sacrificing performance. REACTION consists of two core components: Adaptive Feature Extraction (AFE) leverages mutual information to project high-dimensional sparse features into a compact, informative subspace. This adaptively filters noisy or weak features, reduces embedding parameters, and preserves implicit feature interactions without explicit high-order computation. Dynamic Tower Fusion (DTF) bridges the representational gap between dual-tower expressiveness and single-tower efficiency. It facilitates rich cross-tower interactions during training, then merges the towers into a unified, low-latency single tower for inference. Extensive experiments on four large-scale benchmarks demonstrate that REACTION not only outperforms existing methods in accuracy but also achieves a drastic reduction in both model parameters and inference costs, thus establishing a new paradigm for efficient and scalable recommendation systems.

1 Introduction

Recommender systems predict user preferences, aiming to provide relevant content. Precise modeling of complex user-item interactions is essential, making dual-tower architectures a highly effective dominant framework in contemporary recommender systems (Zhang et al. 2021; Pi et al. 2019; Wang et al. 2023b).

Early dual-tower methods (Huang et al. 2013) used independent user or item embeddings with dot-product matching but lacked fine-grained interactions. Subsequent work (Zhou et al. 2018, 2019) introduced attention and sequential modules to model dynamic interests, while SIM (Pi et al. 2019) and FinalMLP (Mao et al. 2023) boosted expressivity via two-stage retrieval and residual MLP blocks. Furthermore, Transformer-based dual towers equipped with contrastive or ranking losses (Wang et al. 2021) have directly optimized embeddings for retrieval, marking a shift from simple lookups to multi-objective, high-capacity architectures that advance both accuracy and scalability.

Despite their strong predictive performance, contemporary feature-rich recommendation models suffer from significant inefficiencies caused by both feature and structural redundancy. Our analysis reveals that approximately 46% of features contribute negligibly, exhibiting noise-like behavior. Removing these redundant features preserves or even enhances model accuracy (Figure 1(a)–(b)), demonstrating substantial redundancy within the feature set. To mitigate this redundancy, numerous feature selection methods have been proposed. However, many existing approaches (Wang et al. 2020; Ma et al. 2021; Peng, Sugiyama, and Mine 2024; Yang et al. 2023) fail to adequately capture essential dependencies—both among features and between features and labels—dependencies that become critically important under conditions of extreme data sparsity. Furthermore, feature selection techniques (Zhao, Anand, and Wang 2019; Chiew et al. 2019) are susceptible to multicollinearity: highly intercorrelated feature groups are often eliminated entirely during selection, discarding potentially informative predictors. Methods based on mutual information (Zhou, Wang, and Zhu 2022; Beraha et al. 2019) or specialized neural architectures (Wang et al. 2022) can also incur super-quadratic computational complexity (worse than $O(n^2)$), hindering scalability. Concurrently, alternative approaches (Wang et al. 2025; Liu et al. 2020; Guo et al. 2021; Yan et al. 2022) compromise discriminative power by restricting explicit feature interactions to third-order or lower. This limitation curtails model expressiveness, diminishing their capacity to uncover nuanced user-item relationships, particularly in sparse data settings. Therefore, effectively addressing feature redundancy while preserving discriminative power and computational efficiency remains a critical and multifaceted challenge in building high-performance recommender systems for sparse data.

Moreover, feature-level waste tends to amplify structural

*corresponding author

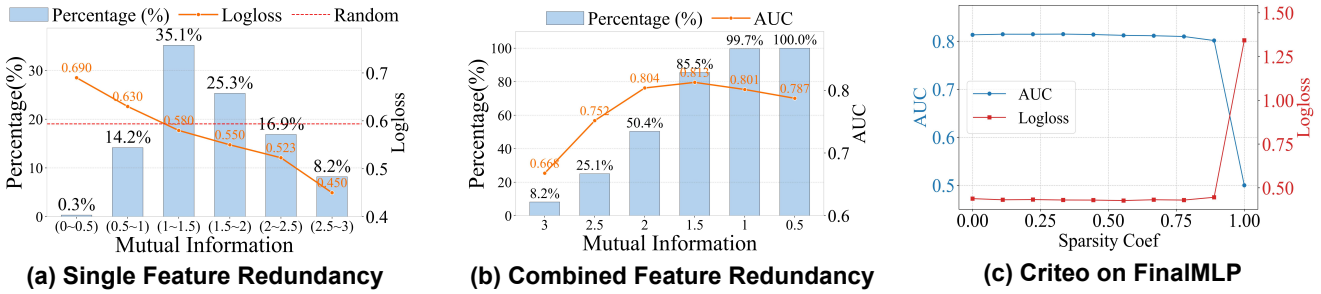

Figure 1: (a)–(b) validate feature redundancy, while (c) target structural redundancy. In (a), features are grouped by mutual information; orange lines indicate group performance, and blue bars show their proportions. "Random" denotes the baseline with randomly initialized features. (b) reports the impact of low-variance features by retaining only those with variance $\leq x$. In (c), we vary the dense layer's sparsity to reveal structural redundancy.

inefficiencies: to compensate for degraded signals, models are often deepened, introducing new computational bottlenecks. Structural redundancy therefore emerges as a parallel concern. While deeper models (Lian et al. 2018; Wu et al. 2023; Zheng et al. 2022; Jia et al. 2025) pursue higher-order patterns, they incur prohibitive computational costs, hindering real-time deployment. Dual-tower architectures (Lang et al. 2021; Huang, Zhang, and Zhang 2019; Wang, She, and Zhang 2021), designed for complex interactions, suffer from inefficient inference due to cross-tower computations and potential representation misalignment. Although knowledge distillation (Deng et al. 2023; Tian et al. 2023; Zhu et al. 2020; Li et al. 2019; Deng et al. 2023) offers a path to efficiency, it typically relies on large teacher models and seldom yields superior student performance. Crucially, Figure 1(c) demonstrate that retaining only 10% of dense-layer parameters preserves 98.5% performance on Criteo, while just 20% recovers 99.7% on Avazu, highlighting the severe structural waste inherent in current designs. This highlights the enduring expressiveness–efficiency dilemma: dual-tower models deliver rich interactions but incur high latency, whereas distilled single-tower models cut latency at the cost of representational capacity.

To overcome the dual challenges of feature redundancy and structural redundancy in modern recommenders, we propose REACTION (paRameter-Efficient leArning for recommendation). REACTION introduces a unified information-theoretic framework with two synergistic modules: Adaptive Feature Extraction (AFE) and Dynamic Tower Fusion (DTF).

AFE tackles feature redundancy by framing feature selection as a mutual information maximization problem. Instead of costly explicit feature crossing or indiscriminate embedding, AFE captures high-order feature interactions through a low-order decomposition of mutual information. It decomposes the complex mutual information into three interpretable and tractable components: (i) Preference Information – a feature's direct relevance to the target prediction, (ii) Behavior Information – its discriminative power for contrasting user behaviors, and (iii) Collaborative Information – its redundancy or synergy relative to already-selected features. This weighted combination of low-order terms enables linear-time selection while implicitly preserving high-order signals. The resulting sparse feature set substantially reduces embedding table size, filters out noisy signals at the source, and enhances the efficiency and quality of downstream representations.

Leveraging this compact feature embedding, DTF redesigns the dual-tower paradigm to resolve structural redundancy. During training, DTF interleaves lightweight adaptive layers between towers, facilitating adaptive bidirectional information propagation at the feature level. Critically, DTF operates on two levels: within each tower, features are softly aligned to learn fine-grained complementary patterns, and across towers, structural relationships are jointly calibrated to promote a shared, structurally aligned representation space. A contrastive objective further refines this process by attracting interaction vectors from the same user-item pair in latent space while repelling those from different pairs – explicitly optimizing for interaction distinctiveness. The key innovation lies in DTF's inherent fusion capability: the architecture actively unifies the tower representations during the training process itself. Specifically, the introduced fusion losses—spatial, relational, and contrastive—drive both towers to learn a shared, semantically aligned representation space. As a result, after training, either tower alone suffices for inference without a significant performance drop. This enables a seamless transition to single-tower inference, where the dual-tower expressiveness is retained through representational unification, while latency and memory costs are significantly reduced.

Experimental results on four large-scale benchmark datasets demonstrate that REACTION consistently outperforms state-of-the-art recommendation models, validating both its architectural effectiveness and strong generalization capability. Our main contributions are as follows:

- We identify significant feature and structural redundancy in modern recommendation models, where most embedding and dense-layer parameters contribute negligibly to performance.
- We propose the Adaptive Feature Extraction module, which reduces feature-selection complexity to linear, and leverages low-order mutual information factors to capture

high-order feature interactions, thereby markedly lowering feature redundancy.

- We introduce the Dynamic Tower Fusion module, which merges dual-tower architecture into a single inference tower, substantially simplifying the model and cutting structural redundancy.
- Extensive experiments on four benchmarks demonstrate that REACTION achieves significant performance gains with high efficiency, validating its effectiveness and robustness.

2 Related Work

2.1 Feature Selection in Recommendation

Efficient recommender systems rely on selecting informative features while minimizing parameter overhead (Du et al. 2024b,a; Jia et al. 2024). Existing methods such as COLD (Wang et al. 2020) and FSCD (Ma et al. 2021) emphasize individual feature relevance but often neglect inter-feature dependencies that are crucial for modeling user behavior. Model-agnostic and permutation-based approaches (Gao et al. 2023) offer broader applicability, yet they suffer from static selection strategies, high computational costs, and an inability to capture synergistic feature interactions. On the other hand, techniques that explicitly model feature interactions (Wang et al. 2025; Liu et al. 2020; Guo et al. 2021; Yan et al. 2022) are typically limited to third-order interactions, whereas capturing high-order interactions is essential for fully understanding complex user-item relationships. Our AFE module approximates high-order mutual information via low-order decomposition, efficiently capturing feature–label relevance, behavior dependencies, and feature synergies. This reduces complexity from exponential to linear, enabling scalable and accurate feature selection while preserving key interdependencies.

2.2 Dual-tower Structure in Recommendation

As deep learning reshaped recommender systems, Deep Crossing first modeled feature interactions via network depth but struggled with high-dimensional sparse data. The Wide&Deep framework (Cheng et al. 2016) became foundational by combining a linear wide part for low-order patterns with a deep network for nonlinear modeling, yet the wide component failed to capture complex high-order dependencies (Cheng et al. 2016; Yang et al. 2022; Sun et al. 2021). Later models(Shan et al. 2016; Guo et al. 2017; Pan et al. 2018) improved expressiveness by introducing CrossNet and FM modules, automating multi-order interaction learning and reducing feature engineering. However, redundant feature interactions introduced noise and overhead. AFM added attention to weight interactions, but ignored feature-type heterogeneity (Lang et al. 2021). FiBiNET(Huang, Zhang, and Zhang 2019) addressed this via bilinear and importance modeling, while (Wang, She, and Zhang 2021; Wang et al. 2023a; Song et al. 2019) further refined selection using gating and residual units. Yet, prior work overlooked unifying dual-tower expressiveness into a single-tower structure at inference time, limiting their ability to balance accuracy with efficiency(Wu et al. 2023; Zheng et al. 2022). Still, most prior work overlooked compressing dual-tower expressiveness into a single-tower structure for inference, limiting their balance between accuracy and real-time efficiency (Liu et al. 2024; Wu et al. 2023; Zheng et al. 2022). Our Dynamic Tower Fusion (DTF) integrates dual-tower capabilities into a unified inference tower using spatial, relational, and contrastive fusion losses, enabling rich, task-aligned representations with minimal computational overhead.

3 Preliminary

Definition of Mutual Information. Mutual information (MI) is a fundamental concept in information theory that quantifies the dependency between two random variables. For random variables X and Y, the mutual information $I(X;Y)$ is defined as:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1}$$

where $H(X)$ and $H(Y)$ denote the entropy of X and Y, respectively, measuring their individual uncertainties, and $H(X,Y)$ is the joint entropy, reflecting the uncertainty of X and Y considered together. Equivalently, MI can be expressed using conditional entropy:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{2}$$

where $H(X|Y)$ and $H(Y|X)$ represent the residual uncertainties of X and Y when the other variable is known. In recommendation, the mutual information $I(S';L)$ between a feature subset S' and labels L serves as the core objective for feature selection, as it directly measures the predictive relevance of S' to user behaviors.

Problem Context and Mathematical Formulation. Given a full feature set $X = \{X_1, X_2, \ldots, X_d\}$ (e.g., user demographics, item attributes, historical interactions) and a label set $L = \{l_1, l_2, \ldots, l_m\}$ (e.g., click and purchase), the goal of feature selection is to find a subset $S' \subseteq X$ ($|S'| = k \ll d$) such that:

$$S' = \arg \max_{S' \subseteq X,\ |S'|=k} I(S';L), \tag{3}$$

where $I(S';L)$ is the mutual information between S' and L, defined as:

$$I(S';L) = H(S') - H(S' \mid L), \tag{4}$$

where $H(S')$ is the entropy of the selected feature subset, measuring the uncertainty in S' and $H(S' \mid L)$ is the conditional entropy of S' given L, measuring the residual uncertainty in S' after observing L.

Maximizing $I(S';L)$ directly targets retaining maximal predictive information while reducing feature dimensionality. However, this optimization is NP-hard due to the combinatorial search over 2^d possible subsets. To address this, we adopt a greedy sequential forward selection (SFS) strategy, iteratively adding features to S' by maximizing the conditional mutual information:

$$J(X_m) = I(X_m; L \mid S), \tag{5}$$

where S is the current selected subset, and X_m is a candidate feature not in S. Here, S denotes the set of already-selected features, and X_m represents a new candidate feature under evaluation.

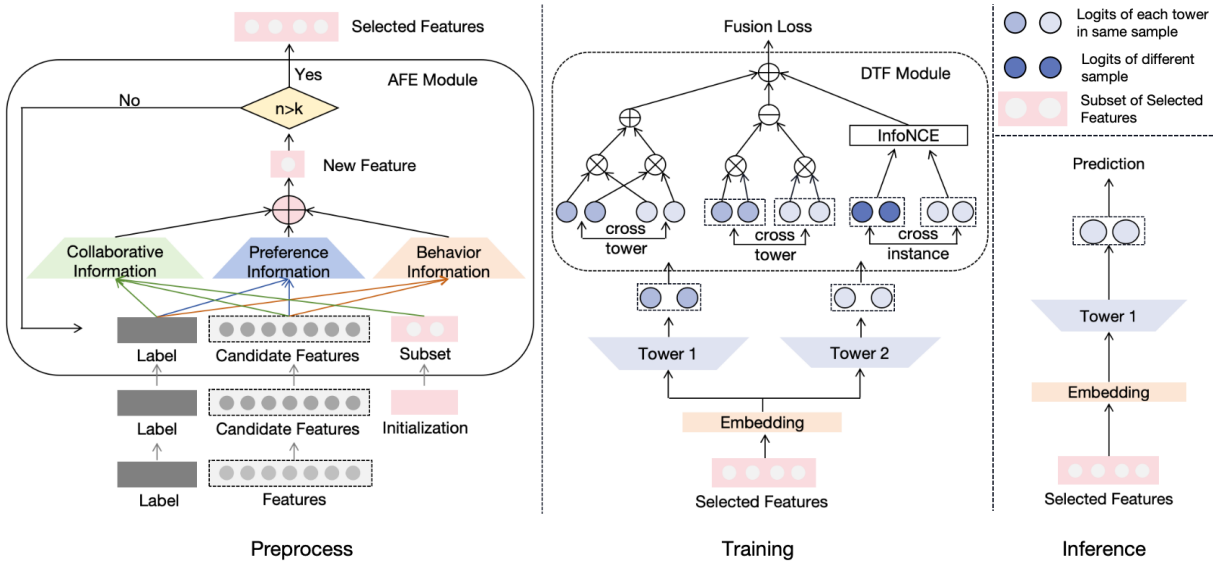

Figure 2: REACTION exploits parameter-efficient learning with AFE and DTF for recommendation.

4 Method

In this section, we introduce REACTION (paRameter-Efficient LeArning for recommendaTION), a unified framework targeting two key challenges: feature redundancy and structural redundancy. As shown in Figure 2, REACTION consists of two main modules: Adaptive Feature Extraction (AFE), which distills low-order informative signals from high-dimensional sparse features, and Dynamic Tower Fusion (DTF), which enables efficient feature interaction through dual collaborative towers. We detail each component below with formal definitions and practical insights.

4.1 Adaptive Feature Extraction (AFE)

The AFE module is designed to reduce both inference latency and overfitting by selecting a small yet highly informative subset of features. The selected subset ensures preservation of the majority of mutual information at three levels: (i) within the feature set, (ii) between features and labels, and (iii) among the labels themselves. Directly evaluating the conditional mutual information:

$$J(X_m) = \sum_{x_m,\, l,\, s} p(x_m, l, s)\, \log \frac{p(l \mid x_m, s)}{p(l \mid s)} \tag{6}$$

requires estimation of the full joint distribution $p(L \mid X_m, S)$, whose complexity grows exponentially in the size of S. This renders naïve computation infeasible for industrial-scale recommendation.

To overcome this barrier, we draw on empirical findings in recommendation that (i) high-order label dependencies are negligible due to extreme sparsity (often more than 95% of user–item interactions receive no positive feedback) and (ii) the bulk of predictive power arises from simple pairwise correlations between individual features and positive feedback outcomes. Guided by these insights, we introduce the Recommendation Label Independence Distribution (RLID) assumption, which approximates

$$p(L \mid X_m, S) \approx \prod_{l_i \in L} \prod_{X_j \in S} p\big(l_i \mid X_m, X_j\big)^{\frac{1}{|L||S|}}, \tag{7}$$

Under RLID, the overwhelmingly skewed marginal $p(l_i)$ toward the opposite labels outcome effectively nullifies the contribution of high-order label combinations, while the influence of a new feature X_m on any label l_i is carried almost entirely through its pairwise interactions with existing features $X_j \in S$.

The exponent $1/(|L|\,|S|)$ both normalizes the product into a valid probability distribution and preserves the marginal conditional statistics of the original joint. As a result, the computational burden of feature selection drops from exponential $O(2^{|L|}\,|S|^{|L|})$ to linear $O(|L|\,|S|)$, enabling efficient, large-scale application with negligible loss of mutual information.

Low-Order Factor Decomposition. To make the mutual information term $I(X_m; L \mid S)$ computationally feasible, we derive a principled approximation that decomposes it into a sum of three interpretable low-order components. This decomposition captures the primary relevance to individual labels, the interactions among labels, and the synergy with already-selected features.

Theorem 1 (Low-Order Factor Decomposition)**.** *Under the RLID assumption, the conditional mutual information between a candidate feature X_m and the label vector L given the selected subset S can be approximated by three low-order*

components:

$$
\begin{aligned}
&I(X_m; L \mid S) \\
&\approx \underbrace{\frac{1}{|L|}\sum_{l_i \in L} I(X_m; l_i)}_{\text{PI}} + \underbrace{\frac{1}{|L|}\sum_{l_i \in L}\sum_{l_j \neq l_i} I\big(X_m; l_j \mid l_i\big)}_{\text{BI}} \\
&+ \underbrace{\frac{1}{|L|\cdot|S|}\sum_{l_i \in L}\sum_{X_j \in S} I\big(X_m; l_i \mid X_j\big)}_{\text{CI}}. \qquad (8)
\end{aligned}
$$

The above decomposition introduces three distinct components, each capturing a different aspect of the relationship between the candidate feature, labels, and selected features. We detail the intuition behind each term below. The preference information(PI) directly uses mutual information $I(X_m; l_i)$ to quantify how informative feature X_m is about each individual user action (e.g. click or purchase), thus capturing its correlation with user preference; behavior information(BI) leverages conditional mutual information $I(X_m; l_j \mid l_i)$ to measure how X_m modulates the dependency (i.e. coupling information) between successive behaviors; and collaborative information(CI) employs $I(X_m; l_i \mid X_j)$ to assess the extra information gain arising from the synergy between the new feature X_m and each already-selected feature X_j, enriching the joint semantic and behavioral signal in the recommendation model.

In this way, the conditional mutual information that originally required estimating high-order distributions is decomposed into three categories of low-order mutual information terms. This achieves efficient feature selection with linear complexity while preserving most of the predictive information. These low-order factors collectively approximate the high-order mutual information while avoiding exponential computational costs. By selecting features that maximize $J(X_m) = \text{PI} + \text{BI} + \text{CI}$, AFE ensures that S' retains the most informative dependencies, mitigating overfitting to noise.

4.2 Dynamic Tower Fusion

Dual-tower models capture complementary biases but double inference cost. Our Dynamic Tower Fusion (DTF) uses multi-level fusion losses during training to transfer and consolidate representations between towers, allowing deployment of just one tower at inference. This preserves dual-tower expressiveness and slashes computational overhead.

Inter-Tower Multi-level Representation Fusion We propose two fusion strategies to enforce consistency between feature maps from the two towers: spatial feature fusion and relational feature fusion.

Given an input processed by towers S_1 and S_2, we obtain feature maps $\mathbf{Z} \in \mathbb{R}^{h\times w\times d_z}$ (from S_1) and $\mathbf{F} \in \mathbb{R}^{h\times w\times d_f}$ (from S_2). To enable fusion, a learnable projection $W \in \mathbb{R}^{d_f \times d_z}$ is applied to $\mathbf{Z}$, producing $\mathbf{Z}' = W\mathbf{Z} \in \mathbb{R}^{h\times w\times d_f}$, matching $\mathbf{F}$ in spatial and feature dimensions.

To fuse features spatially, we minimize the spatial fusion loss $\mathcal{L}_{\text{srf}}$, which encourages corresponding local features to be similar. For each spatial location (i, j), we compute the cosine similarity between $\mathbf{z}'_{ij}$ and $\mathbf{f}_{ij}$, and define:

$$
\mathcal{L}_{\text{srf}} = \frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w}\text{ReLU}\left(1 - m_1 - \frac{\mathbf{z}'_{ij}\cdot\mathbf{f}_{ij}}{\|\mathbf{z}'_{ij}\|\|\mathbf{f}_{ij}\|}\right), \qquad (9)
$$

where m_1 is a margin that reduces the penalty for well-aligned feature pairs, focusing training on harder-to-align cases.

To fuse relational information, we propose the relational fusion loss $\mathcal{L}_{\text{rrf}}$, which aligns the pairwise similarity structures within each feature map. Denoting $N = h \times w$ as the number of spatial features, and $\mathbf{z}_i, \mathbf{z}_j$ and $\mathbf{f}_i, \mathbf{f}_j$ as spatial features from $\mathbf{Z}'$ and $\mathbf{F}$ respectively, we enforce consistency between their intra-sample similarity matrices:

$$
\mathcal{L}_{\text{rrf}} = \frac{1}{N^2}\sum_{i,j=1}^{N}\text{ReLU}\left(\left|\frac{\mathbf{z}_i^\top\mathbf{z}_j}{\|\mathbf{z}_i\|\|\mathbf{z}_j\|} - \frac{\mathbf{f}_i^\top\mathbf{f}_j}{\|\mathbf{f}_i\|\|\mathbf{f}_j\|}\right| - m_2\right), \qquad (10)
$$

where m_2 is a relaxation margin. This loss preserves the geometric relationships between features, ensuring the structural patterns in $\mathbf{Z}'$ and $\mathbf{F}$ remain consistent.

Cross-Tower Instance-Contrastive Fusion To further enhance sample-level discriminative fusion, we introduce a cross-tower contrastive loss based on InfoNCE. For a mini-batch of size B, let $\mathbf{z}_i^{(1)}$ and $\mathbf{z}_i^{(2)} \in \mathbb{R}^d$ be L2-normalized hidden features of the same sample from towers S_1 and S_2. Positive pairs are $(\mathbf{z}_i^{(1)}, \mathbf{z}_i^{(2)})$, while other pairs are negatives.

The contrastive loss for sample i is defined as:

$$
\ell_i = -\log\frac{\exp(\mathbf{z}_i^{(1)\top}\mathbf{z}_i^{(2)}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{z}_i^{(1)\top}\mathbf{z}_j^{(2)}/\tau)}, \qquad (11)
$$

$$
\ell'_i = -\log\frac{\exp(\mathbf{z}_i^{(2)\top}\mathbf{z}_i^{(1)}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{z}_i^{(2)\top}\mathbf{z}_j^{(1)}/\tau)}. \qquad (12)
$$

The final cross-tower contrastive fusion loss is:

$$
\mathcal{L}_{\text{contra}} = \frac{1}{2B}\sum_{i=1}^{B}(\ell_i + \ell'_i). \qquad (13)
$$

This loss maximizes a lower bound on the mutual information between tower representations, aligning their semantics while avoiding trivial collapse.

Joint Optimization Objective

The overall loss for each tower $S_k \in \{S_1, S_2\}$ is:

$$
\mathcal{L}_{S_k} = \mathcal{L}_{\text{CE}}(y, \hat{y}_{S_k}) + \alpha\mathcal{L}_{\text{srf}} + \beta\mathcal{L}_{\text{rrf}} + \gamma\mathcal{L}_{\text{contra}}, \qquad (14)
$$

where $\mathcal{L}_{\text{CE}}$ is the supervised cross-entropy loss, and α, β, γ are hyperparameters balancing each term.

5 Experiments

5.1 Experimental Setup

Dataset We evaluate REACTION on four public benchmarks: Criteo and Avazu (Cheng and Xue 2021; Zhu et al. 2023), MovieLens (Harper and Konstan 2015) and UGC (Wu et al. 2023). Table 2 summaries the statistics of the datasets.

Model	Criteo		Avazu		MovieLens		UGC				
	AUC ↑	Logloss ↓	AUC ↑	Logloss ↓	AUC ↑	Logloss ↓	AUC ↑	Logloss ↓	Params ↓	FLOPs ↓	Latency ↓
FmFM	0.8112	0.4408	0.7744	0.3831	0.9464	0.2095	0.7948	0.4176	8.42M	17.82M	0.204 ms
FwFM	0.8104	0.4414	0.7741	0.3835	0.9415	0.2151	0.7937	0.4103	9.11M	16.27M	0.218 ms
xDeepFM	0.8122	0.4407	0.7821	0.3799	0.9513	0.2044	0.8000	0.4104	329.96M	3921.64M	0.821 ms
DCNV2	0.8127	0.4394	0.7838	0.3782	0.9546	0.2029	0.8023	0.4114	110.83M	124.74M	0.472 ms
FiBiNet	0.8126	0.4415	0.7837	0.3783	0.9460	0.2091	0.7999	0.4107	742.74M	795.49M	0.458 ms
AutoInt+	0.8126	0.4456	0.7832	0.3786	0.9537	0.2088	0.8010	0.4127	56.28M	85.19M	0.486 ms
FiGNN	0.8109	0.4412	0.7830	0.3799	0.9539	0.2032	0.7998	0.4106	19.29M	73.38M	0.912 ms
ECKD	0.8123	0.4422	0.7834	0.3838	0.9551	0.1979	0.8026	0.4108	51.25M	52.52M	0.863 ms
BKD	0.8125	0.4408	0.7845	0.3822	0.9556	0.1983	0.8029	0.4048	40.63M	48.37M	0.538 ms
APG	0.8130	0.4402	0.7846	0.3825	0.9555	0.1976	0.8016	0.4066	38.13M	42.66M	0.513 ms
KD-DAGFM	0.8132	0.4390	0.7852	0.3780	0.9566	0.1989	0.8031	0.4059	7.21M	14.41M	0.172 ms
AutoDis	0.8141	0.4381	0.7859	0.3788	0.9562	0.1980	0.8030	0.4055	12.35M	18.41M	0.374 ms
FinalMLP	0.8139	0.4380	0.7855	0.3784	0.9561	0.1976	0.8034	0.4044	62.41M	87.28M	0.556 ms
LLM-ESR	0.8143	0.4376	0.7863	0.3780	0.9569	0.1982	0.8042	0.4049	135.87M	184.37M	0.983 ms
REACTION	**0.8165****	**0.4358****	**0.7882****	**0.3761****	**0.9590****	**0.1956****	**0.8055****	**0.4018****	**4.83M**	**7.58M**	**0.148 ms**

Table 1: Overall performance comparison in the four datasets. The t-test results show that our performance advantage over the previous SOTA method is statistically significant (Liu et al. 2019) ($p < 10^{-2}$). $\star$: $p < 10^{-2}$, $\star\star$: $p < 10^{-4}$.

Baseline and Evaluation Metrics In highlighting the versatility of our methodology, we exclusively deploy two distinct sizes of tiny MLPs as two towers, setting the sizes to [80, 80, 80] and [240] as our respective baselines. In the assessment of model performance, several key metrics are employed to provide a comprehensive understanding of their effectiveness, including: AUC (Area Under the Curve) (Zhou et al. 2018), Logloss(Logarithmic Loss) (Xu et al. 2021), Params (Parameters) (Xu and Wu 2020), FLOPs (Floating Point Operations Per Second) and Latency. Numerous studies (Mao et al. 2023; Chen et al. 2021) highlight that even a marginal improvement at the 0.001-level in AUC can yield substantial revenue gains for a company, particularly when dealing with a sizable user base.

Datasets	# Features	# Fields	# Instances
Criteo	2.1M	39	45M
Avazu	1.5M	23	40M
MovieLens	90K	3	2M
UGC	10.6M	92	102M

Table 2: Statistics of the evaluation datasets.

Experimental Settings To ensure a fair comparison, we run each method for 10 times with different random seeds on a single GPU (NVIDIA GeForce RTX 3090), and the average testing performance is reported. Subsequently, two-tailed t-tests (Liu et al. 2019) are conducted to compare the performance. We optimize all the models with mini-batch Adam (Kingma 2014), where the learning rate is searched from$\{10e-6, 5e-5, ..., 10e-2\}$. The weight of $\mathbf{L_2}$ regularization is tuned in $\{10e-6, 5e-5, ..., 10e-3\}$. We reuse the baseline models and apply grid search to find the optimal settings. The feature embedding dimension is set to 16, and the learning rate ranges from 1e-3, 1e-4, to 1e-5. Utilizing Adam (Jiang et al. 2023) as the optimizer. Distillation loss weights, denoted as α_2 and β_2, exhibit variations within the intervals of 1e-1, 1, 10 and 10, 100, 1000, respectively. For KD-DAGFM, the number of propagation modules is examined with varying values of 2, 3, to 4. For FwFM and FmFM, a field-wise linear weight scheme is adopted. AutoInt+ is configured with a depth of 2, 2 heads, and an attention size of 40. In the case of Ensemble of Cross Knowledge Distillation (ECKD), teachers, including AutoInt+, DCNV2, and xDeepFM, are selected, and a "soft label+pre-train" scheme is applied. The sizes of MLPs in FinalMLP are set to [400,400,400] and [1200].

5.2 Experimental Results

Comparison with the SOTA methods Our REACTION adopts only compact MLPs as towers, yet as shown in Table 1, our model achieves the highest AUC and lowest Logloss while incurring the fewest parameters, minimal FLOPs, and the lowest latency—demonstrating both strong generalization and exceptional efficiency. Notably, deeper models (e.g., xDeepFM, AutoInt+) outperform shallower ones (e.g., FwFM, FmFM), and dual-tower architectures (e.g., DCNV2, FinalMLP) further enhance multi-level feature interactions. However, no baseline consistently excels across all metrics: increased model complexity often brings diminishing returns in performance while substantially raising computational costs. Meanwhile, distillation-based methods (ECKD, BKD, KD-DAGFM) accelerate inference but rely on costly multi-teacher training. In contrast, our single-stage architecture not only outperforms all baseline methods but also strikes a compelling balance between representational capacity and computational efficiency.

Model	Criteo AUC ↑	Avazu AUC ↑	Movielens AUC ↑	UGC AUC ↑
Baseline	0.8118	0.7833	0.9543	0.8022
Baseline + ER	0.8122	0.7835	0.9549	0.8027
Baseline + KD	0.8124	0.7836	0.9553	0.8024
Baseline + DTF	0.8154	0.7872	0.9579	0.8040

Table 3: Comparison of AUC performance for different models and datasets.

Model	DFF	AUC ↑	Logloss ↓
All Fields	None	0.8008	0.4196
PCA	[1, 5, 11, 20, 21, 28]	0.7988	0.4211
LASSO	[2, 20, 22, 28, 39]	0.7993	0.4218
IDARTS	Unstable	0.7962	0.4234
AutoField	[1, 12, 22, 49, 60, 65]	0.8012	0.4184
TayFCS	[5,34,47,52,63,65]	0.8023	0.4125
Baseline+AFE	[5, 47, 57, 60, 63, 65]	0.8030	0.4062

Table 4: Comparison of AFE and variants on the UGC dataset. DFF denotes dropped feature fields.

5.3 Ablation Studies

Comparison of embedding modules and feature interaction modules We compare dynamic tower fusion (DTF) with its variants such as entropy regularization (ER) and knowledge distillation (KD) for comparison. The result is shown in Table 3. We can see that the DTF module ouperforms ER and KD across the four public datasets, demonstrating its strong generalization ability. Additionally, We also compare our adaptive feature extraction (AFE) method with representative methods based on our "Baseline", including PCA (Wold, Esbensen, and Geladi 1987), LASSO (Tibshirani 1996), IDARTS (Jiang et al. 2019), AutoField (Wang et al. 2022) and TayFCS (Wang et al. 2025) in Table 4. The results show that AFE offers greater stability, higher AUC, and lower Logloss in all the four datasets. We evaluated various thresholds for these feature selection methods and reported the best performance for each method at different thresholds.

T/S	Baseline	ECKD	KD-DAGFM	Baseline+DTF	Rel. Imp
1	0.8115	0.8123	0.8132	**0.8154**	0.0022
2	0.8121	0.8132	0.8142	**0.8167**	0.0025
3	0.8135	0.8140	0.8149	**0.8178**	0.0029
4	0.8138	0.8146	0.8156	**0.8187**	0.0031
5	0.8140	0.8154	0.8160	**0.8197**	0.0037
6	0.8142	0.8165	0.8163	**0.8205**	0.0042
7	0.8144	0.8170	0.8166	**0.8211**	0.0041
8	0.8145	0.8173	0.8168	**0.8212**	0.0039

Table 5: AUC performance of various models with differing numbers of teachers/towers on the Criteo dataset.

T/S	Baseline	ECKD	KD-DAGFM	Baseline+DTF	Rel. Imp
1	0.8024	0.8026	0.8031	**0.8057**	0.0026
2	0.8033	0.8036	0.8039	**0.8070**	0.0031
3	0.8038	0.8044	0.8046	**0.8081**	0.0035
4	0.8041	0.8050	0.8053	**0.8090**	0.0037
5	0.8044	0.8055	0.8056	**0.8098**	0.0042
6	0.8046	0.8060	0.8059	**0.8106**	0.0046
7	0.8050	0.8062	0.8061	**0.8113**	0.0051
8	0.8052	0.8063	0.8065	**0.8115**	0.0050

Table 6: AUC performance of various models with differing numbers of teachers/towers on the UGC dataset.

Scaling REACTION with Larger Tower Cohorts Previous experiments employed cohorts of 2 towers. This section investigates REACTION's scalability with larger cohorts, presenting results on Avazu and Criteo (Tables 5, 6). As cohort size increases in "Baseline + DTF", the average performance of individual towers improves, widening the gap versus independently trained models. This demonstrates that tower generalization is significantly enhanced by learning within progressively larger groups. Performance improves overall with more towers. However, the marginal gain per additional tower exhibits an inverted U-shape: initially increasing, then diminishing. Crucially, new towers should introduce architectural or parametric diversity distinct from existing ones, as each tower encodes unique prior information. Similar new towers risk biasing learning towards a subset, limiting overall efficacy. Extensive experimentation confirms that greater tower diversity typically yields better model performance.

6 Conclusion

In this work, we address two forms of redundancy in recommender systems—feature and structural—via two lightweight, model-agnostic modules. AFE mitigates feature redundancy through adaptive selection based on low-order mutual information terms, while DTF resolves structural redundancy by enabling collaborative learning among towers. When integrated into the REACTION framework, these modules yield state-of-the-art accuracy and substantially reduce parameter overhead. Our findings provide actionable insights for both academic research and industrial deployment, while highlighting the importance of future exploration into generalized parameter-efficient learning for recommender systems.

References

Beraha, M.; Metelli, A. M.; Papini, M.; Tirinzoni, A.; and Restelli, M. 2019. Feature selection via mutual information: New theoretical insights. In *2019 international joint conference on neural networks (IJCNN)*, 1–9. IEEE.

Chen, B.; Wang, Y.; Liu, Z.; Tang, R.; Guo, W.; Zheng, H.; Yao, W.; Zhang, M.; and He, X. 2021. Enhancing explicit and implicit feature interactions via information sharing for parallel deep ctr models. In *Proceedings of the 30th ACM international conference on information & knowledge management*, 3757–3766.

Cheng, H.-T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, 7–10.

Cheng, Y.; and Xue, Y. 2021. Looking at ctr prediction again: Is attention all you need? In *Proceedings of the 44th International ACM SIGIR conference on research and development in information retrieval*, 1279–1287.

Chiew, K. L.; Tan, C. L.; Wong, K.; Yong, K. S.; and Tiong, W. K. 2019. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences*, 484: 153–166.

Deng, Y.; Chen, Y.; Dong, X.; Pan, L.; Li, H.; Cheng, L.; and Mo, L. 2023. BKD: A Bridge-based Knowledge Distillation Method for Click-Through Rate Prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1859–1863.

Du, Z.; Chen, J.; Jia, Q.; Wu, C.; Zhu, J.; Dong, Z.; and Tang, R. 2024a. LightCS: Selecting Quadratic Feature Crosses in Linear Complexity. In *Companion Proceedings of the ACM Web Conference 2024*, 38–46.

Du, Z.; Wu, C.; Jia, Q.; Zhu, J.; and Chen, X. 2024b. A Tutorial on Feature Interpretation in Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 1281–1282.

Gao, W.; Hao, P.; Wu, Y.; and Zhang, P. 2023. A unified low-order information-theoretic feature selection framework for multi-label learning. *Pattern Recognition*, 134: 109111.

Guo, H.; Chen, B.; Tang, R.; Zhang, W.; Li, Z.; and He, X. 2021. An embedding learning framework for numerical features in ctr prediction. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2910–2918.

Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247*.

Harper, F. M.; and Konstan, J. A. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4): 1–19.

Huang, P.-S.; He, X.; Gao, J.; Deng, L.; Acero, A.; and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2333–2338.

Huang, T.; Zhang, Z.; and Zhang, J. 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.

Jia, P.; Du, Z.; Wang, Y.; Zhao, X.; Li, X.; Wang, Y.; Liu, Q.; Guo, H.; and Tang, R. 2025. SELF: Surrogate-light Feature Selection with Large Language Models in Deep Recommender Systems. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, 1145–1155.

Jia, P.; Wang, Y.; Du, Z.; Zhao, X.; Wang, Y.; Chen, B.; Wang, W.; Guo, H.; and Tang, R. 2024. Erase: Benchmarking feature selection methods for deep recommender systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5194–5205.

Jiang, J.; Zhang, P.; Luo, Y.; Li, C.; Kim, J. B.; Zhang, K.; Wang, S.; Xie, X.; and Kim, S. 2023. AdaMCT: adaptive mixture of CNN-transformer for sequential recommendation. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, 976–986.

Jiang, Y.; Hu, C.; Xiao, T.; Zhang, C.; and Zhu, J. 2019. Improved differentiable architecture search for language modeling and named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3585–3590.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lang, L.; Zhu, Z.; Liu, X.; Zhao, J.; Xu, J.; and Shan, M. 2021. Architecture and operation adaptive network for online recommendations. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 3139–3149.

Li, Z.; Cui, Z.; Wu, S.; Zhang, X.; and Wang, L. 2019. Fi-gnn: Modeling feature interactions via graph neural networks for ctr prediction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 539–548.

Lian, J.; Zhou, X.; Zhang, F.; Chen, Z.; Xie, X.; and Sun, G. 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.

Liu, B.; Tang, R.; Chen, Y.; Yu, J.; Guo, H.; and Zhang, Y. 2019. Feature generation by convolutional neural network for click-through rate prediction. In *The World Wide Web Conference*, 1119–1129.

Liu, B.; Zhu, C.; Li, G.; Zhang, W.; Lai, J.; Tang, R.; He, X.; Li, Z.; and Yu, Y. 2020. Autofis: Automatic feature interaction selection in factorization models for click-through rate prediction. In *proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2636–2645.

Liu, Q.; Wu, X.; Wang, Y.; Zhang, Z.; Tian, F.; Zheng, Y.; and Zhao, X. 2024. Llm-esr: Large language models enhancement for long-tailed sequential recommendation. *Advances in Neural Information Processing Systems*, 37: 26701–26727.

Ma, X.; Wang, P.; Zhao, H.; Liu, S.; Zhao, C.; Lin, W.; Lee, K.-C.; Xu, J.; and Zheng, B. 2021. Towards a Better Trade-off between Effectiveness and Efficiency in Pre-Ranking: A Learnable Feature Selection based Approach. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2036–2040.

Mao, K.; Zhu, J.; Su, L.; Cai, G.; Li, Y.; and Dong, Z. 2023. FinalMLP: An Enhanced Two-Stream MLP Model for CTR Prediction. *arXiv preprint arXiv:2304.00902*.

Pan, J.; Xu, J.; Ruiz, A. L.; Zhao, W.; Pan, S.; Sun, Y.; and Lu, Q. 2018. Field-weighted factorization machines for click-through rate prediction in display advertising. In *Proceedings of the 2018 World Wide Web Conference*, 1349–1357.

Peng, S.; Sugiyama, K.; and Mine, T. 2024. Less is more: Removing redundancy of graph convolutional networks for recommendation. *ACM Transactions on Information Systems*, 42(3): 1–26.

Pi, Q.; Bian, W.; Zhou, G.; Zhu, X.; and Gai, K. 2019. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2671–2679.

Shan, Y.; Hoens, T. R.; Jiao, J.; Wang, H.; Yu, D.; and Mao, J. 2016. Deep crossing: Web-scale modeling without manually crafted combinatorial features. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 255–262.

Song, W.; Shi, C.; Xiao, Z.; Duan, Z.; Xu, Y.; Zhang, M.; and Tang, J. 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.

Sun, Y.; Pan, J.; Zhang, A.; and Flores, A. 2021. FM2: Field-matrixed factorization machines for recommender systems. In *Proceedings of the Web Conference 2021*, 2828–2837.

Tian, Z.; Bai, T.; Zhang, Z.; Xu, Z.; Lin, K.; Wen, J.-R.; and Zhao, W. X. 2023. Directed Acyclic Graph Factorization Machines for CTR Prediction via Knowledge Distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 715–723.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1): 267–288.

Wang, F.; Gu, H.; Li, D.; Lu, T.; Zhang, P.; and Gu, N. 2023a. Towards Deeper, Lighter and Interpretable Cross Network for CTR Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2523–2533.

Wang, R.; Shivanna, R.; Cheng, D.; Jain, S.; Lin, D.; Hong, L.; and Chi, E. 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.

Wang, X.; Du, Z.; Zhu, J.; Wu, C.; Jia, Q.; and Dong, Z. 2025. TayFCS: Towards Light Feature Combination Selection for Deep Recommender Systems. *arXiv preprint arXiv:2507.03895*.

Wang, Y.; Du, Z.; Zhao, X.; Chen, B.; Guo, H.; Tang, R.; and Dong, Z. 2023b. Single-shot feature selection for multi-task recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 341–351.

Wang, Y.; Zhao, X.; Xu, T.; and Wu, X. 2022. Autofield: Automating feature selection in deep recommender systems. In *Proceedings of the ACM Web Conference 2022*, 1977–1986.

Wang, Z.; She, Q.; and Zhang, J. 2021. Masknet: Introducing feature-wise multiplication to CTR ranking models by instance-guided mask. *arXiv preprint arXiv:2102.07619*.

Wang, Z.; Zhao, L.; Jiang, B.; Zhou, G.; Zhu, X.; and Gai, K. 2020. Cold: Towards the next generation of pre-ranking system. *arXiv preprint arXiv:2007.16122*.

Wold, S.; Esbensen, K.; and Geladi, P. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3): 37–52.

Wu, S.-L.; Du, L.; Yang, J.-Q.; Wang, Y.-A.; Zhan, D.-C.; Zhao, S.; and Sun, Z.-X. 2023. Re-sort: Removing spurious correlation in multilevel interaction for ctr prediction. *arXiv preprint arXiv:2309.14891*.

Xu, C.; and Wu, M. 2020. Learning feature interactions with lorentzian factorization machine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6470–6477.

Xu, Y.; Zhu, Y.; Yu, F.; Liu, Q.; and Wu, S. 2021. Disentangled self-attentive neural networks for click-through rate prediction. In *Proceedings of the 30th ACM international conference on information & knowledge management*, 3553–3557.

Yan, B.; Wang, P.; Zhang, K.; Li, F.; Deng, H.; Xu, J.; and Zheng, B. 2022. Apg: Adaptive parameter generation network for click-through rate prediction. *Advances in Neural Information Processing Systems*, 35: 24740–24752.

Yang, L.; Wang, S.; Tao, Y.; Sun, J.; Liu, X.; Yu, P. S.; and Wang, T. 2023. Dgrec: Graph neural network for recommendation with diversified embedding generation. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, 661–669.

Yang, X.; Peng, X.; Wei, P.; Liu, S.; Wang, L.; and Zheng, B. 2022. Adasparse: Learning adaptively sparse structures for multi-domain click-through rate prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4635–4639.

Zhang, W.; Qin, J.; Guo, W.; Tang, R.; and He, X. 2021. Deep learning for click-through rate estimation. *arXiv preprint arXiv:2104.10584*.

Zhao, Z.; Anand, R.; and Wang, M. 2019. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. In *2019 IEEE international conference on data science and advanced analytics (DSAA)*, 442–452. IEEE.

Zheng, J.; Chen, S.; Du, Y.; and Song, P. 2022. A multiview graph collaborative filtering by incorporating homogeneous and heterogeneous signals. *Information Processing & Management*, 59(6): 103072.

Zhou, G.; Mou, N.; Fan, Y.; Pi, Q.; Bian, W.; Zhou, C.; Zhu, X.; and Gai, K. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5941–5948.

Zhou, G.; Zhu, X.; Song, C.; Fan, Y.; Zhu, H.; Ma, X.; Yan, Y.; Jin, J.; Li, H.; and Gai, K. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1059–1068.

Zhou, H.; Wang, X.; and Zhu, R. 2022. Feature selection based on mutual information with correlation coefficient. *Applied intelligence*, 52(5): 5457–5474.

Zhu, J.; Jia, Q.; Cai, G.; Dai, Q.; Li, J.; Dong, Z.; Tang, R.; and Zhang, R. 2023. Final: Factorized interaction layer for ctr prediction. In *Proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval*, 2006–2010.

Zhu, J.; Liu, J.; Li, W.; Lai, J.; He, X.; Chen, L.; and Zheng, Z. 2020. Ensembled CTR prediction via knowledge distillation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2941–2958.