HIT Model: A Hierarchical Interaction-Enhanced Two-Tower Model for Pre-Ranking Systems

Haoqiang Yang* Tencent Shenzhen, Guangdong, China yanghaoqiang1998@gmail.com

Mengzhuo Guo[†] Sichuan University Chengdu, Sichuan, China mengzhguo@scu.edu.cn Congde Yuan*
Tencent
Shenzhen, Guangdong, China
congdeyuan@gmail.com

Wei Yang Tencent Shenzhen, Guangdong, China viviwyang@tencent.com Kun Bai Tencent Shenzhen, Guangdong, China bookerbai@tencent.com

Chao Zhou Tencent Shenzhen, Guangdong, China derekczhou@tencent.com

ABSTRACT

Online display advertising platforms rely on pre-ranking systems to efficiently filter and prioritize candidate ads from large corpora, balancing relevance to users with strict computational constraints. The prevailing two-tower architecture, though highly efficient due to its decoupled design and pre-caching, suffers from cross-domain interaction and coarse similarity metrics, undermining its capacity to model complex user-ad relationships. In this study, we propose the Hierarchical Interaction-Enhanced Two-Tower (HIT) model, a new architecture that augments the two-tower paradigm with two key components: generators that pre-generate holistic vectors incorporating coarse-grained user-ad interactions through a dual-generator framework with a cosine-similarity-based generation loss as the training objective, and multi-head representers that project embeddings into multiple latent subspaces to capture fine-grained, multifaceted user interests and multi-dimensional ad attributes. This design enhances modeling effectiveness without compromising inference efficiency. Extensive experiments on public datasets and largescale online A/B testing on Tencent's advertising platform demonstrate that HIT significantly outperforms several baselines in relevance metrics, yielding a 1.66% increase in Gross Merchandise Volume and a 1.55% improvement in Return on Investment, alongside similar serving latency to the vanilla two-tower models. The HIT model has been successfully deployed in Tencent's online display advertising system, serving billions of impressions daily. The code is available at https://anonymous.4open.science/r/HIT model-5C23.

CCS CONCEPTS

• Information systems → Computational advertising; Display advertising; Novelty in information retrieval.

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

Conference acronym 'XX, June 03-05, 2018, Woodstock, NY

KEYWORDS

Online display advertising, Pre-ranking, Two-tower model, Deep learning

ACM Reference Format:

1 INTRODUCTION

Online display advertising systems serve as a critical revenue stream for numerous digital platforms, including Google [2, 24], Meta [22, 28], Alibaba [8, 13, 21], and Tencent [4, 26, 27, 30]. These systems recommend the most interesting advertisements to users and facilitate them to purchase the products, thereby optimizing advertiser profits [4, 29]. Typically organized in a cascaded architecture, as depicted in Figure 1, these systems commence with *targeting*, which identifies a candidate set of ad impressions from a vast corpus comprising millions of ads. This candidate set is subsequently *scored* to distill several hundred most relevant ads. Finally, the *ranking* and *re-ranking* stages refine this selection to a handful of ads, completing the pipeline.

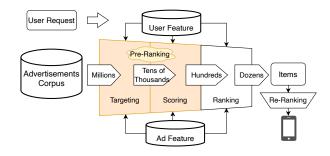


Figure 1: Online display advertising system.

The *pre-ranking* system, encompassing the targeting and scoring tasks highlighted in the shaded area of Figure 1, plays a crucial role in the architecture of online display advertising systems. This phase is critical as it precedes the evaluation and final display of ads [23]. Its primary objective is to ensure that the ad impressions are

^{*}Both authors contributed equally to this research.

 $^{^{\}dagger}$ Corresponding author.

not only relevant to the targeted users, thereby enhancing *model* effectiveness, but also that they are processed in a manner that is not computationally burdensome, ensuring inference efficiency. Consequently, the principal challenge lies in achieving an optimal balance between these two tasks [10].

Due to the stringent requirement that pre-ranking systems in large-scale advertising platforms must filter millions of ad candidates per user within milliseconds, the standard *two-tower model* has become the de facto in industry [5, 9, 19]. While this architecture excels in computational efficiency by pre-caching techniques, its effectiveness is constrained by (1) a lack of coarse-grained information exchange with the opposite tower, preventing cross-domain dependency modeling, and (2) a simple dot-product-based similarity computation, failing to capture finegrained semantic relationships between user and ad embeddings. Therefore, traditional two-tower-based models may miss the critical multi-faceted user interests and multi-dimensional ad attributes essential for precise matching.

To address these two limitations, we propose a Hierarchical Interaction-Enhanced Two-Tower (HIT) model, as illustrated in Figure 3. The proposed HIT model is comprised of two new components: the *generators* and *multi-head representers*. The generators establish a holistic representation of user/ad static features, enabling the early generation of embeddings to mimic the coarsegrained information from the counterpart tower before the multihead representers. To this end, we adopt a cosine-similarity-based generation loss and separately learn positive and negative samples by two generators. Such a dual-generator approach treats positive and negative samples differently since they contain rich information about the targeted and non-targeted user/ad.

The multi-head representers employ linear projections to map the generated user/ad embeddings into different latent sub-spaces. To effectively extract user/ad representations, we first capture the strongest signal from each vector of the user's multi-faceted interest (or the ad's multi-dimensional attributes) to determine the most relevant user/ad interactions, and then aggregate these interactions to obtain a final score. It accounts for the fine-grained multi-faceted user interests and multi-dimensional ad attributes, yielding more precise matching scores. Moreover, the multi-head representers retain the advantage of pre-cached ad embeddings, thereby maintaining high serving efficiency.

Our contributions can be summarized as follows. First, we propose the HIT model for pre-ranking systems, which addresses the issue of insufficient information interaction while preserving the serving efficiency of the two-tower paradigm. The HIT model contains two new components: the generators for pre-generating coarse-grained vectors of user/ad cross-domain feature fusions and multi-head representers for projecting embeddings into multifaceted user interests (or multi-dimensional ad attributes). Second, we find that using a single vector to embed user/ad characteristics, as stated in traditional two-tower models, is insufficient to capture user/ad granular characteristics. The proposed generators employ a cosine-similarity-based loss to eliminate vector magnitude discrepancies for more accurate coarse-grained generation vectors, while the multi-head representers adopt a max-then-sum operation to accomplish fine-grained matching between multi-faceted user interests and multi-dimensional ad attributes. Finally, comprehensive

evaluations of public datasets demonstrate that the HIT model consistently outperforms baseline models. Rigorous online A/B tests on Tencent's online display advertising platform further confirm its effectiveness, with significant gains in GMV, ROI, and reduced response times. The HIT model has been successfully deployed on a large-scale advertising platform, impacting billions of daily impressions.

2 RELATED WORK

The two-tower architecture has emerged as a dominant paradigm in pre-ranking systems, widely employed in online display advertising and recommender systems. As pre-ranking systems operate at the forefront of the decision pipeline, their efficiency and effectiveness significantly impact subsequent stages. The two-tower model leverages two independent neural networks to encode users and ads, thereby enabling joint training while also allowing each tower to be pre-trained offline to minimize computational overhead during online inference. This design not only ensures scalability but also achieves lower latency in prediction compared to traditional shallow learning methods, such as logistic regression [12] and gradient boosting decision trees [18]. Consequently, the two-tower model has become a cornerstone in modern recommender and advertising systems [5, 10]. We categorize it into four mainstream approaches based on their interaction mechanisms: vanilla, early interaction, late interaction, and all-to-all interaction models, as illustrated in Figure 2.

Vanilla Two-tower Model. The vanilla two-tower model, exemplified by DSSM [5], encodes user and ad features independently without explicit interaction between the two sides. This simplicity enables efficient offline graph construction and nearest-neighbor retrieval, ensuring low-latency online inference. However, as the complexity of user and ad features increases, the absence of explicit feature interactions limits the model's ability to fully capture underlying relationships, resulting in suboptimal performance compared to more advanced methods.

Early Interaction Model. This strategy involves feature interaction after the embedding layers but before the encoding modules (Figure 2b). For instance, Yu et al. [25] proposes a Dual Augmented Two-tower model (DAT) that learns auxiliary vectors for users and ads based on ads and user embedding vectors, respectively. Such a process is similar to "distillation" that keeps the information most appropriate for describing the user-ad interactions [9, 11, 19]. Another recent research, the mixture of virtual-kernel experts (MVKE), has modeled user-ad interactions through a multi-task learning scheme to learn user interests in different ads [23].

Late Interaction Model. Instead of modeling feature interactions before the encoding module, the late interaction directly impacts the encoded users and ads vectors, as shown in Figure 2c. Compared to early interaction models, this type incorporates deeper interactions between users and ads since it works after transforming the embedded vectors through shallow Deep Neural Networks (DNNs) [14]. For example, to accelerate the computational efficiency, Humeau et al. [7] developed the poly-encoder to learn cross-features through a final attention mechanism structure. Li et al. [10] propose a lightweight Multi-Layer Perceptron (MLP),

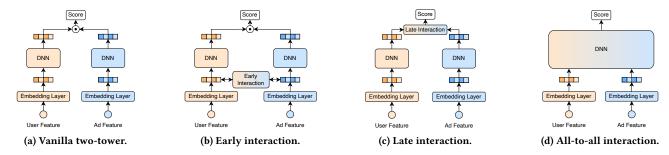


Figure 2: Illustrations of four types of two-tower model structures.

termed the IntTower, which is introduced to model complex cross features between users and ads.

All-to-all Interaction Model. The all-to-all interaction utilizes an MLP to directly model the user-ad feature interactions right after the embedding layers (Figure 2d). This approach departs from the paradigm that separately encodes the embedded user and ad vectors, thus overcoming the limitations of inner product representations in two-tower models. In this way, the advanced model structures, such as Wide & Deep [1], Deep & Cross Networks (DCN) [17], AutoInt [15], and COLD [20], can be used to account for high-order and extremely complex user-ad interactions, thus enhancing the model accuracy. However, ad vectors cannot be pre-computed and stored with all-to-all interaction, significantly slowing inference speed.

In summary, vanilla two-tower models are pioneers in balancing inference efficiency and computational effectiveness. Early and late interaction models extend the two-tower framework to enhance feature interactions, improving model effectiveness. In contrast, all-to-all interaction models completely abandon the two-tower structure for more thorough feature interactions, sacrificing inference efficiency for greater model effectiveness. Our proposed HIT model retains the two-tower structure, establishing latent correlations between users and ads, and fully leveraging the rich and diverse characteristics of users and ads. It better aligns with real-world data distributions, resulting in significant performance improvements that surpass those of all-to-all interaction models without compromising inference efficiency.

3 PRELIMINARIES

3.1 Problem Definition

In online display advertising, when a user request is received, the system combines user features \mathbf{x}_u (static feature \mathbf{x}_{ub} : age, gender, occupation; dynamic feature \mathbf{x}_{ug} : location, active period) and ad features \mathbf{x}_a (static feature \mathbf{x}_{ab} : category, tag; dynamic feature \mathbf{x}_{ag} : click count) to compute scores for all eligible ads via $s(\mathbf{x}_u, \mathbf{x}_a)$, where $s(\cdot)$ is a DNN trained on historical data to predict user click probabilities. A top-k truncation step then selects the most user-preferred ads based on these scores.

3.2 Basic Structure

The basic structure of a vanilla two-tower model, as shown in the "two-tower backbone" part (in green) in Figure 3, consists of two distinct components: a user tower and an ad tower, along with a scoring and online serving module.

3.2.1 User Tower. The user features \mathbf{x}_{ub} and \mathbf{x}_{ug} are first passed through an embedding layer, converting the sparse features into dense vectors $\mathbf{e}_{ub} \in \mathbb{R}^{d \times n_{ub}}$ and $\mathbf{e}_{ug} \in \mathbb{R}^{d \times n_{ug}}$, where d denotes the embedding dimension, n_{ub} and n_{ug} are the numbers of user static and dynamic features, respectively. These dense vectors are subsequently concatenated to obtain $\mathbf{e}_{u} = [\mathbf{e}_{ub}, \mathbf{e}_{ug}] \in \mathbb{R}^{d \times (n_{ub} + n_{ug})}$, which is then fed into a DNN with L layers:

$$\mathbf{h}^{(l)} = \text{ReLu}(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}), \text{ for } l = 1, 2, \dots, L,$$
 (1)

where $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$, $\mathbf{b}^{(l)} \in \mathbb{R}^{d_l}$, $\mathbf{h}^{(l)} \in \mathbb{R}^{d_l}$ are the weight matrix, bias vector, and output of the l-th layer, with d_l being the layer width, and $\mathbf{h}^{(0)} = \mathbf{e}_u$, $d_0 = d \times (n_{ub} + n_{ug})$. The final user embedding is obtained through L_2 norm: $\mathbf{h}_u = \|\mathbf{h}^{(L)}\|_2$.

3.2.2 Ad Tower. Similarly to the user tower, the ad features \mathbf{x}_{ab} and \mathbf{x}_{ag} are converted into $\mathbf{e}_{ab} \in \mathbb{R}^{d \times n_{ab}}$ and $\mathbf{e}_{ag} \in \mathbb{R}^{d \times n_{ag}}$, where n_{ab} and n_{ag} denote the numbers of ad static and dynamic features. These vectors are concatenated to form $\mathbf{e}_a = [\mathbf{e}_{ab}, \mathbf{e}_{ag}] \in \mathbb{R}^{d \times (n_{ab} + n_{ag})}$, and are fed into a DNN. The final ad embedding \mathbf{h}_a is obtained by applying L_2 norm to the output of the last layer.

3.2.3 Combining Embedded User and Ad Vectors. After the user embedding and ad embedding are computed, a score \hat{y} is obtained by inner product:

$$\hat{\mathbf{y}} = \mathbf{h}_{u}^{T} \mathbf{h}_{a}. \tag{2}$$

In the pre-ranking system, each training sample is constituted in the form of $(\mathbf{x}_{ub}, \mathbf{x}_{ug}, \mathbf{x}_{ab}, \mathbf{x}_{ag}, y)$ with $y \in \{0, 1\}$ serving as the label, where 1 indicates a positive signal (the user likes the ad), and 0 otherwise. The model parameters can be optimized by minimizing the discrepancy between output \hat{y} and the label y.

In online systems, ad embeddings are pre-computed and cached for efficient retrieval [5], therefore only the requested user embedding need to be calculated on the fly, which significantly improves computational efficiency. However, the lack of user-ad feature interaction leads to inaccurate predictions due to missing information in the construction of \mathbf{h}_u and \mathbf{h}_a . Thus, we propose the HIT model.

4 THE PROPOSED HIT MODEL

The overall architecture of the proposed HIT model, depicted in Figure 3, extends the vanilla two-tower model through the incorporation of two novel modules: *coarse-grained generation* (in pink) and *fine-grained matching* (in purple).

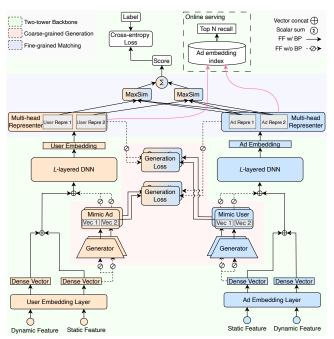


Figure 3: Overview of the HIT model architecture.

4.1 Coarse-grained Generation

The key component of coarse-grained generation is the **generator**, which is a simple multilayer perceptron network $g_k(\cdot)$ with two hidden layers that accepts *only* static features $(\mathbf{x}_{ub} \text{ or } \mathbf{x}_{ab})$ as input and produces representations $\mathbf{r}_k^{\mathbf{m}} \in \mathbb{R}^p$ manifesting information about the opposite tower. For example, the computation process of the user tower can be expressed as follows:

$$\mathbf{r}_{k,\delta}^{\mathbf{m}} = g_k(\mathbf{x}_{\delta b}), \quad \text{for } k = 1, 2, \dots, K,$$
 (3)

where $\mathbf{r}_{k,\delta}^{\mathbf{m}}$ denotes the mimic users/ads (with $\delta = u/a$ representing user and ad, respectively; the subscript δ in $\mathbf{x}_{\delta b}$ carries the same meaning) vectors $\mathbf{r}^{\mathbf{m}}$ in generator k, and K is a hyperparameter indicating the number of generators, note that K is related to the number of label types. For notation simplicity, we use $\mathbf{r}_k^{\mathbf{m}}$ as a substitute for $\mathbf{r}_{k,\delta}^{\mathbf{m}}$, focusing on the user side. In this study, since the label has positive and negative types, we set K=2. Nevertheless, we can use smaller K by feeding only positive or negative samples to the generator and greater K by defining more types of labels. The training process of a generator involves comparing its output with those produced by the following multi-head representer, from which a generation loss is computed. The generation loss will be elaborated on in the following subsection.

It is noteworthy that generators simulate the high-level representations of the opposing tower, i.e., the generators in the user tower simulate ad representations, while the generators in the ad tower simulate user representations. The motivation can be intuitively explained by an example. Suppose a wealthy young woman(user) is more likely to be interested in expensive luxury brands (target ads) than cheap, low-quality goods (non-target ads). The generator in the user-tower mimics the the target ads information so that the user-tower is exposed to the ads information. To this end, the static

features, describing the fundamental interests and attributes of the users and ads, are more important for a generator than the dynamic features. Thus, the proposed HIT model uses only static features as input for generators to fit the high-level representations of target and non-target entities, constructing correlations between static features and high-level representations through neural networks.

After obtaining the mimic representations $\mathbf{r}_k^{\mathbf{m}}$, they need to be first normalized to unit length using L2 normalization, followed by a simple concatenation operation $f(\cdot)$ to consolidate them with the dense vector \mathbf{e}_u . Denote the concatenate vector as \mathbf{e}^i :

$$e^{i} = f(r_{1}^{m}, r_{2}^{m}, \dots, r_{K}^{m}, e_{u}).$$
 (4)

which is fed into a DNN, resulting in the user embedding $\mathbf{h}_u^{\mathbf{i}} \in \mathbb{R}^{d_L}$. Similarly, for the ad tower, the ad embedding $\mathbf{h}_a^{\mathbf{i}} \in \mathbb{R}^{d_L}$ is obtained.

4.2 Fine-grained Matching

In this module, the user/ad embedding $\mathbf{h}_u^i/\mathbf{h}_a^i$ are fed into a **multi-head representer** to capture multi-dimensional user-ad interaction. In particular, the multi-head representer maps user/ad embedding to different latent sub-spaces to extract more informative representation, mirroring that a user has multi-faceted interests and an ad has multi-dimensional attributes. The projection process for the user u can be described mathematically as follows:

$$\mathbf{r}_{u,(j)} = \mathbf{W}_{u,(j)} \mathbf{h}_{u}^{i} + \mathbf{b}_{u,(j)}, \quad \text{for } j = 1, 2, \dots, J,$$
 (5)

where J is a hyperparameter indicating the number of sub-spaces mapped by the multi-head representer, $\mathbf{W}_{u,(j)} \in \mathbb{R}^{z \times d_L}$ and $\mathbf{b}_{u,(j)} \in \mathbb{R}^z$ are the transformation matrix and bias vector for the j-head sub-space, $\mathbf{r}_{u,(j)} \in \mathbb{R}^z$ is the high-level representation for user of j-head in the multi-head representer. Similarly, the representation vector $\mathbf{r}_{a,(j)} \in \mathbb{R}^z$ for ad a under j-head sub-space is expressed as:

$$\mathbf{r}_{a,(j)} = \mathbf{W}_{a,(j)} \mathbf{h}_a^{\mathbf{i}} + \mathbf{b}_{a,(j)}, \quad \text{for } j = 1, 2, \dots, J,$$
 (6)

where $\mathbf{W}_{a,(j)} \in \mathbb{R}^{z \times d_L}$ and $\mathbf{b}_{a,(j)} \in \mathbb{R}^z$ are the transform matrix and bias vector.

We consider the matching degree between each user interest and all ad attributes in order to discover the most appropriate matching ad attribute, and then sum up all facets of interest. In this way, the final score \hat{y} is obtained by:

$$\hat{y} = \sum_{i_{-}=1}^{J} \max_{j_{a} \in \{1, 2, \dots, J\}} \left\{ (\mathbf{r}_{u, (j_{u})})^{T} \mathbf{r}_{a, (j_{a})} \right\}.$$
 (7)

In online serving, ad representations $\mathbf{r}_{a,(j)}$ are pre-computed and cached in advance, thereby retaining the computational efficiency advantage of the two-tower architecture.

4.3 Model Optimization

As illustrated in Figure 3, the optimization of the HIT model involves two types of losses: the generation loss and the cross-entropy loss.

4.3.1 Generation Loss. The J head transformations $\mathbf{r}_{u,(j)}$ obtained from Eq.(5) can be concatenated for parallel computing, and the result is the user representation vector $\mathbf{r}_u = [\mathbf{r}_{u,(1)}, \mathbf{r}_{u,(2)}, \dots, \mathbf{r}_{u,(J)}] \in \mathbb{R}^{z \times J}$ (the same process to get $\mathbf{r}_a \in \mathbb{R}^{z \times J}$). The generation loss \mathcal{L}_{qu} ,

which is designed to minimize the distance between \mathbf{r}_u and $\mathbf{r}_k^{\mathbf{m}} \in \mathbb{R}^p$ derived from Eq.(3), where $p = z \times J$, is represented as follows:

$$\mathcal{L}_{gu} = -\frac{1}{N} \sum_{i=1}^{N} (y_i \text{Dist}(\mathbf{r}_u, \mathbf{r}_1^{\mathbf{m}}) + (1 - y_i) \text{Dist}(\mathbf{r}_u, \mathbf{r}_2^{\mathbf{m}})), \quad (8)$$

where N is the total number of samples, y_i is the ground truth label of i-th sample, and $\mathrm{Dist}(\cdot)$ is cosine distance. Cosine distance is preferred here because it focuses on vector orientation (reflecting user/ad characteristics) rather than magnitude, aligning with the generator's objective of distinguishing distinct representations. Generator 1 ($\mathbf{r}_1^{\mathbf{m}}$, target generator) and Generator 2 ($\mathbf{r}_2^{\mathbf{m}}$, non-target generator) are trained on positive and negative samples, respectively. Consequently, \mathcal{L}_{gu} aggregates these distances across all generators in the user tower.

The generation loss \mathcal{L}_{ga} for the ad tower can be computed in an identical procedure as in Eq.(8).

4.3.2 Cross-entropy Loss. We use cross-entropy loss (CEloss) to calculate the difference between the predicted scores and the true labels, which is defined as follows:

$$\mathcal{L}_{c} = -\frac{1}{N} \sum_{i=1}^{N} (y_{i} \log(\sigma(\hat{y}_{i})) + (1 - y_{i}) \log(1 - \sigma(\hat{y}_{i}))), \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid function , and \hat{y} is the *i*-th prediction score derived from Eq.(7).

4.3.3 Total Loss. The total loss \mathcal{L} is calculated as follows:

$$\mathcal{L} = \mathcal{L}_c + \alpha (\mathcal{L}_{au} + \mathcal{L}_{aa}), \tag{10}$$

where α is a hyperparameter used to balance the weight of the cross-entropy loss and generation loss, as shown in Figure 3, the HIT model is trained using the backpropagation algorithm in an end-to-end manner. However, to enable generators to focus on fitting high-level representations without interfering with the backbone network in the model, both the input and output of generators undergo "stop gradient" operations. Additionally, to ensure that the generation loss does not affect the training itself, the representation outputs of the multi-head representer are subjected to "stop gradient" operations within the computation of the generation loss.

5 EXPERIMENTS

In this section, we answer the following questions:

- (Q1) What are the HIT model's performances compared to baselines?
- (Q2) What is the impact of each component on the performance?
- (O3) How to explain the interaction mechanism in the HIT model?
- (Q4) What are the online performances of the HIT model?

5.1 Experimental Setup

In this section, three public datasets, MovieLens, Amazon (Electro), and Alibaba, are used as offline evaluation datasets. Following Huang et al. [6], the samples are randomly divided into two parts: 80% for training and the remaining 20% for testing. This ensures that our model is trained on a sufficiently large dataset while still having enough samples to evaluate its performance on unknown data. The detailed statistics of datasets are provided in Table 1.

We empirically set the feature embedding dimension to 32 and the training batch size to 256. Each DNN consists of a three-layer

Dataset	Users	Items	Samples
MovieLens-1M	6,040	3,952	1,000,000
Amazon(Electro)	192,403	630,001	1,689,188
Alibaba	1,061,768	785,597	26,557,961

Table 1: Basic statistics of each dataset.

MLP with hidden dimensions [300, 300, 32]. The parameters of the HIT model are temporarily fixed as: $\alpha=10^{-3}$ in Eq.(10); J=2 in Eq.(5) and Eq.(6); output dimension z=16 for the multi-head representer. The generator is a two-layered MLP with [64, 32]. Other detailed implementations can be found in the open-source code. We conduct experiments with 2 Tesla T4 GPUs.

5.2 Experimental Results Compared to Baselines

Table 2 answers *Q1* and reports the average results when comparing the proposed HIT model to the following baselines: **DSSM** [5], **DAT** [25], **MVKE** [23], **Poly-Encoder** [7], **IntTower** [10], **Wide&Deep** [1], **DeepFM** [3], **DCN** [17], **AutoInt** [15], and **COLD** [20]. These models are introduced in Section 2 and can be classified into a specific two-tower structure. We consider three metrics: **CEloss**, as shown in Eq.(9); **Area Under the Curve** (AUC), the area under the Receiver Operating Characteristic curve; and **Relative Improvement**, measuring the relative improvement in AUC [31].

Compared to baseline models, we observe a performance hierarchy across user-ad interaction types: vanilla < early interaction < late interaction < all-to-all interaction. While the vanilla two-tower model is computationally efficient, its limited user-ad interaction leads to significantly lower AUC, underscoring its representational limitations. Early- and late-interaction models (e.g., DAT, MVKE, IntTower) offer improvements but still fall short in capturing the multi-faceted nature of user interests and ad attributes. All-to-all interaction models demonstrate stronger performance due to their end-to-end interaction modeling; however, they lack pre-computability, resulting in substantial inference overhead and limited industrial applicability.

The proposed HIT model outperforms all baselines across datasets in both AUC and CEloss, indicating superior discriminative ability for user-ad matching. This advantage arises from two key components: (1) the generator, which effectively bridges static features and high-level embeddings, enabling HIT to surpass even complex models such as DCN and AutoInt with lower computational cost; and (2) the multi-head representer, which captures diverse user and ad characteristics through fine-grained subspace projections, thereby enhancing overall matching precision.

5.3 Ablation Study

To answer Q2, we conduct several ablation experiments to demonstrate the impact of generators, multi-head representers, and their settings on the HIT model.

Q2.1. What is the impact of removing generators or multihead representers? We remove either the generators, the multihead representers, or both, to validate their impacts on the HIT model. Panel A in Table 3 reports the performances. The results show that removing either component, generators or representers,

Туре	Model	# params.	Alibaba		MovieLens		Amazon				
			AUC	CEloss	RelaImpr	AUC	CEloss	RelaImpr	AUC	CEloss	RelaImpr
Vanilla	DSSM [5]	0.74M	0.6579	0.2292	0%	0.8697	0.4559	0%	0.8469	0.4313	0%
Early	DAT [25]	0.76M	0.6598	0.2279	1.20%	0.8712	0.4556	0.40%	0.8480	0.4278	0.31%
	MVKE [23]	0.84M	0.6625	0.2276	2.91%	0.8720	0.4511	0.62%	0.8503	0.4324	0.98%
Late	Poly-Encoder [7]	0.87M	0.6657	0.2273	4.94%	0.8734	0.3971	1.00%	0.8595	0.3855	3.63%
	IntTower [10]	1.18M	0.6827	0.2245	15.71%	0.8974	0.3128	7.49%	0.8696	0.3309	7.91%
All-to-all	Wide&Deep [1]	0.68M	0.6814	0.2250	14.88%	0.8820	0.3344	3.32%	0.8615	0.3409	4.20%
	DeepFM [3]	0.68M	0.6820	0.2247	15.26%	0.8920	0.3211	6.03%	0.8643	0.3405	5.05%
	DCN [17]	0.68M	0.6831	0.2244	15.96%	0.8964	0.3151	7.22%	0.8665	0.3366	5.65%
	AutoInt [15]	0.70M	0.6867	0.2238	18.24%	0.8948	0.3192	6.79%	0.8686	0.3351	6.25%
	COLD [20]	0.68M	0.6816	0.2248	15.01%	0.8836	0.3297	3.75%	0.8633	0.3402	4.72%
Hierarchical	HIT(ours)	0.79M	0.7226	0.2104	40.98%	0.9048	0.3026	9.49%	0.8784	0.3225	9.08%

Table 2: Average results compared to baselines over 10 repetitions. The comparison baseline for the RelaImpr metric is the vanilla two-tower model. (RelaImpr = $\left(\frac{\text{AUC}(\text{model})-0.5}{\text{AUC}(\text{base})-0.5}-1\right) \times 100\%$.) The improvement is significant at $\alpha=0.01$.

	Alibaba		MovieLens		Amazon		
Model	AUC	CEloss	AUC	CEloss	AUC	CEloss	
HIT	0.7226	0.2104	0.9048	0.3026	0.8784	0.3225	
Panel A: Remove generators and multi-head representers							
w/o generators	0.7170	0.2147	0.8955	0.3155	0.8690	0.3325	
w/o representers	0.7166	0.2159	0.8954	0.3157	0.8689	0.3326	
w/o both	0.6579	0.2292	0.8697	0.4559	0.8469	0.4313	
Panel B: Remove target and non-target generators							
w/o target	0.7196	0.2127	0.9022	0.3129	0.8751	0.3251	
w/o non-target	0.7216	0.213	0.8987	0.3144	0.8723	0.3286	
w/o both	0.7170	0.2147	0.8955	0.3155	0.8690	0.3325	
Panel C: Consider dynamic features							
w/ dynamic	0.7205	0.2638	0.9012	0.3107	0.8711	0.3321	
feature	0.7203	0.2036	0.9012	0.3107	0.6/11	0.3321	
Panel D: Consider other types of generation loss							
w/ MAE	0.7204	0.2674	0.8945	0.3316	0.8691	0.3324	
w/ MSE	0.7188	0.265	0.8933	0.3194	0.8689	0.3323	

Table 3: Main ablation experimental results. Each panel represents a set of ablation studies compared to the HIT model.

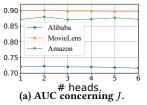
leads to a significant decline in model performance. Removing both components causes the model's accuracy to plummet. Hence, both the generators and the multi-head representers are essential and cannot be omitted.

Q2.2. What is the impact of generator settings on the HIT model? One of the critical settings in the generator is the separation of the target and non-target user/ad information when handling positive and negative samples, which contributes to different informational gains. In Panel B in Table 3, we present how such settings impact the model performances. As expected, the HIT model, including both target and non-target generators, performs the best while the one excluding them performs the worst across all datasets. This indicates that the positive and negative samples have different impacts on generators. However, their contributions can vary depending on the data pattern, as evidenced by the fact that excluding only the target generator performs better on the MovieLens and Amazon datasets. In contrast, excluding only the non-target generator performs better on the Alibaba dataset.

Q2.3. Does dynamic features help increase HIT performances? As shown in Figure 3 and Eq.(3), the generators' input

only accounts for users/ads static features. This helps the generator capture long-term characteristics of users/ads and leave out short-term interest and attribute changes when mimicking the opposite towers. We have tried to include the dynamic features and report the results in Panel C in Table 3. We can observe that including dynamic features decreases the model's performance. This stems from the fact that the dynamic features introduce extra noise when mimicking user and ad representations, thereby introducing erroneous information into the DNN module during feature crossing.

Q2.4. Can other distance metrics replace the cosine similarity in the proposed generation loss? We replace the measurement of the distance in Eqs.(8) by mean square error (MSE) and mean absolute error (MAE). Panel D in Table 3 reports the model performances. We can observe that the HIT model with cosine achieves the best performance on all datasets. Cosine similarity is capable of handling high-dimensional user and ad representations because it addresses the direction rather than the Euclidean distances. In contrast, the metrics MSE and MAE account for the straight distance between two points in space, determining geometric proximity between vectors, so they are more suitable for low-dimensional representations.



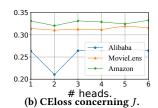


Figure 4: Evaluation metrics concerning *J*.

Q2.5. What is the best number of heads in our multi-head representer? A critical setting for capturing fine-grained semantic relationships between user and ad embeddings is the number of heads *J*. We have tuned different values ranging from 1 to 6 on all datasets, and plotted the curves between the evaluation metrics and *J* in Figure 4. The results show that smaller numbers of heads tend to achieve better performances than greater *J*. The performance

obtains the best when J=2, which is the setting in the proposed HIT model. An excessive number of heads may lead to a complex model structure and suffer the risk of overfitting, which reduces the model's generalization capability. In practice, we recommend a small J, but it can increase given the complexity of the data at hand.

5.4 Investigation on Interaction Mechanism

To answer Q3, we investigate the details of how the proposed interaction mechanism works in practice. We randomly select one user (ad) and all ads (users) related to the selected one in the MovieLens test set to study both the user and ad towers mechanisms.

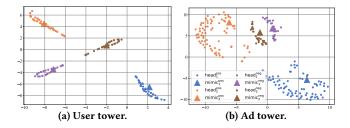


Figure 5: Visualization of mimic/original representations.

The mimic vectors $\mathbf{r}_k^{\mathbf{m}}$ produced by the generator and the multihead representations $\mathbf{r}_{u,(j)}/\mathbf{r}_{a,(j)}$ from the representer are visualized via t-SNE [16] in Figure 5. Circles denote high-dimensional outputs from the representer, while triangles denote mimic vectors generated by the generator.

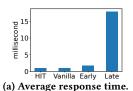
Colors indicate different sample types and attention heads. In Figure 5a, $\mathbf{r}_{u,(1)}$ and $\mathbf{r}_{u,(2)}$ —denoted as $head_1^{pos}$ and $head_2^{pos}$ —represent user-favored (positive) ads, whereas $head_1^{neg}$ and $head_2^{neg}$ correspond to representations of negative samples. The generator's mimic vectors are labeled as $mimic_1^{pos}$, $mimic_2^{pos}$, $mimic_1^{neg}$, and $mimic_2^{neg}$, derived from generators trained on positive and negative samples, respectively. Figure 5b follows a similar notation.

As shown in Figure 5, the multi-head representer forms coherent clusters for semantically similar samples while maintaining clear inter-group separation. This indicates its ability to project $\mathbf{h}_u^{\mathbf{i}}$ and $\mathbf{h}_a^{\mathbf{i}}$ into distinct semantic subspaces, effectively capturing the diverse attributes of users and ads through its multi-perspective representation.

Notably, the generator's mimic vectors $\mathbf{r}_k^{\mathbf{m}}$ tend to lie near the centers of their respective clusters, serving as representative proxies. This centrality suggests that the generator effectively captures the shared characteristics within clusters, indicating successful alignment between static features and high-level embeddings via learned latent correlations.

5.5 Online A/B Test

To address $\it Q4$, we conduct rigorous online A/B testing on a leading commercial display advertising platform. The proposed HIT model is trained on a large-scale dataset comprising over 3.6 billion samples, involving more than 1 billion users and 10 million ad candidates. In the online environment, the HIT model selects several hundred ads per user for forwarding to the downstream modules



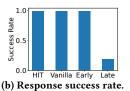


Figure 6: Results of online efficiency. *Millisecond* and *Success Rate* represent the inference time and the success rate of responses under QPS=35,000.

illustrated in Figure 1. These outputs form the treatment group. For the control group, we deploy the MVKE model, which is selected as the strongest-performing baseline on our private data. The two groups' subsequent decision-making processes remain identical to ensure a fair comparison. The A/B test is conducted over five days.

Table 4 provides the improvement of two key business metrics, including GMV and ROI, compared to the control group. There is a 1.55% increase in GMV, indicating that the treatment group brings more revenue for the advertisers than the control group. An increase in ROI means that additional revenue does not require extra costs. The online experimental results demonstrate that the proposed HIT model can better approximate the user-ad pairs and help the following decisions to be more effective.

	GMV	ROI
Improvement	+1.66%	+1.55%

Table 4: Relative improvement compared to control group.

Figure 6 plots the response time and success rate concerning Queries Per Second (QPS) to examine the efficiency. Given the same QPS, the model with a smaller response time and higher success rate is more efficient. We exclude the comparison to all-to-all interaction models since they cannot sustain the QPS=35,000 (A QPS of 35,000 is the threshold metric for online services; if this threshold is not met, online deployment is not feasible). We can see that the HIT model's efficiency is very close to that of the vanilla two-tower models. The efficiency can be summarized as follows: vanilla > early/HIT > late \gg all-to-all models.

6 CONCLUSION

We addressed the *efficient-effectiveness* challenge in pre-ranking systems for online display advertising by introducing the Hierarchical Interaction-enhanced Two-tower (HIT) model. HIT integrates a *generator* that identifies latent correlations between static features and high-level embeddings to enhance coarse-grained user-ad interactions through vector generation, while employing a *multihead representer* that models multi-faceted user/ad attributes via latent subspace projections for fine-grained matching with improved scoring precision. Extensive experiments demonstrate that HIT consistently outperforms state-of-the-art baselines in predictive performance. Real-world deployment on the Tencent advertising platform yielded substantial business gains, including a 1.66% increase in GMV and a 1.55% improvement in ROI, validating HIT's effectiveness and practical scalability in industrial settings.

REFERENCES

- [1] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems. 7–10.
- [2] Jiaping Gui, Stuart Mcilroy, Meiyappan Nagappan, and William GJ Halfond. 2015. Truth in advertising: The hidden cost of mobile ads for software developers. In 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, Vol. 1. IEEE, 100–110.
- [3] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. arXiv preprint arXiv:1703.04247 (2017).
- [4] Mengzhuo Guo, Wuqi Zhang, Congde Yuan, Binfeng Jia, Guoqing Song, Hua Hua, Shuangyang Wang, and Qingpeng Zhang. 2024. A Bayesian Multi-Armed Bandit Algorithm for Bid Shading in Online Display Advertising. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 4506–4513.
- [5] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 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.
- [6] Tongwen Huang, Zhiqi Zhang, and Junlin Zhang. 2019. FiBiNET: combining feature importance and bilinear feature interaction for click-through rate prediction. In Proceedings of the 13th ACM conference on recommender systems. 169–177.
- [7] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. arXiv preprint arXiv:1905.01969 (2019).
- [8] Cong Jiang, Zhongde Chen, Bo Zhang, Yankun Ren, Xin Dong, Lei Cheng, Xinxing Yang, Longfei Li, Jun Zhou, and Linjian Mo. 2024. GATS: Generative Audience Targeting System for Online Advertising. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2920–2924.
- [9] Dan Li, Yang Yang, Hongyin Tang, Jiahao Liu, Qifan Wang, Jingang Wang, Tong Xu, Wei Wu, and Enhong Chen. 2022. VIRT: Improving representation-based text matching via virtual interaction. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 914–925.
- [10] Xiangyang Li, Bo Chen, HuiFeng Guo, Jingjie Li, Chenxu Zhu, Xiang Long, Sujian Li, Yichao Wang, Wei Guo, Longxia Mao, et al. 2022. Inttower: the next generation of two-tower model for pre-ranking system. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 3292–3301.
- [11] Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, et al. 2022. Erniesearch: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval. arXiv preprint arXiv:2205.09153 (2022).
- [12] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. 2013. Ad click prediction: a view from the trenches. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 1222–1230
- [13] Wentao Ouyang, Xiuwu Zhang, Shukui Ren, Li Li, Kun Zhang, Jinmei Luo, Zhaojie Liu, and Yanlong Du. 2021. Learning graph meta embeddings for cold-start ads in click-through rate prediction. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1157–1166.
- [14] Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. 2016. Product-based neural networks for user response prediction. In 2016 IEEE 16th international conference on data mining (ICDM). IEEE, 1149–1154.

- [15] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. Autoint: Automatic feature interaction learning via self-attentive neural networks. In Proceedings of the 28th ACM international conference on information and knowledge management. 1161–1170.
- [16] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. Journal of machine learning research 9, 11 (2008).
- [17] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In Proceedings of the ADKDD'17. 1–7.
- [18] Yaozheng Wang, Dawei Feng, Dongsheng Li, Xinyuan Chen, Yunxiang Zhao, and Xin Niu. 2016. A mobile recommendation system based on logistic regression and gradient boosting decision trees. In 2016 international joint conference on neural networks (TICNN). IEEE, 1896–1902.
- [19] Zekun Wang, Wenhui Wang, Haichao Zhu, Ming Liu, Bing Qin, and Furu Wei. 2021. Distilled dual-encoder model for vision-language understanding. arXiv preprint arXiv:2112.08723 (2021).
- [20] Zhe Wang, Liqin Zhao, Biye Jiang, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. 2020. COLD: Towards the Next Generation of Pre-Ranking System. CoRR abs/2007.16122 (2020). arXiv preprint arXiv:2007.16122 (2020).
- [21] Yanheng Wei, Lianghua Huang, Yanhao Zhang, Yun Zheng, and Pan Pan. 2022. An intelligent advertisement short video production system via multi-modal retrieval. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 3368–3372.
- [22] Melanie Wiese, Carla Martínez-Climent, and Dolores Botella-Carrubi. 2020. A framework for Facebook advertising effectiveness: A behavioral perspective. Journal of Business Research 109 (2020), 76–87.
- [23] Zhenhui Xu, Meng Zhao, Liqun Liu, Lei Xiao, Xiaopeng Zhang, and Bifeng Zhang. 2022. Mixture of virtual-kernel experts for multi-objective user profile modeling. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 4257–4267.
- [24] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In Proceedings of the 13th ACM conference on recommender systems. 269–277.
- [25] Yantao Yu, Weipeng Wang, Zhoutian Feng, and Daiyue Xue. 2021. A dual augmented two-tower model for online large-scale recommendation. DLP-KDD (2021).
- [26] Congde Yuan, Mengzhuo Guo, Chaoneng Xiang, Shuangyang Wang, Guoqing Song, and Qingpeng Zhang. 2022. An actor-critic reinforcement learning model for optimal bidding in online display advertising. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 3604–3613.
- [27] Hengyu Zhang, Junwei Pan, Dapeng Liu, Jie Jiang, and Xiu Li. 2024. Deep Pattern Network for Click-Through Rate Prediction. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1189–1199.
- [28] Wei Zhang, Dai Li, Chen Liang, Fang Zhou, Zhongke Zhang, Xuewei Wang, Ru Li, Yi Zhou, Yaning Huang, Dong Liang, et al. 2024. Scaling User Modeling: Large-scale Online User Representations for Ads Personalization in Meta. In Companion Proceedings of the ACM on Web Conference 2024. 47–55.
- [29] Xiangyu Zhao, Changsheng Gu, Haoshenglun Zhang, Xiwang Yang, Xiaobing Liu, Jiliang Tang, and Hui Liu. 2021. Dear: Deep reinforcement learning for online advertising impression in recommender systems. In Proceedings of the AAAI conference on artificial intelligence, Vol. 35. 750–758.
- [30] Zuowu Zheng, Changwang Zhang, Xiaofeng Gao, and Guihai Chen. 2022. HIEN: hierarchical intention embedding network for click-through rate prediction. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 322–331.
- [31] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 1059–1068.