SaviorRec: Semantic-Behavior Alignment for Cold-Start Recommendation

Yining Yao* yaoyining.yyn@taobao.com Alibaba Group Hangzhou, China

Boya Du boya.dby@taobao.com Alibaba Group Hangzhou, China Ziwei Li* jinhe.lzw@taobao.com Alibaba Group Hangzhou, China

Jialin Zhu xiafei.zjl@taobao.com Alibaba Group Hangzhou, China Shuwen Xiao shuwen.xsw@alibaba-inc.com Alibaba Group Hangzhou, China

Junjun Zheng fangcheng.zjj@alibaba-inc.com Alibaba Group Hangzhou, China

Xiangheng Kong yongheng.kxh@alibaba-inc.com Alibaba Group Hangzhou, China

Abstract

In recommendation systems, predicting Click-Through Rate (CTR) is crucial for accurately matching users with items. To improve recommendation performance for cold-start and long-tail items, recent studies focus on leveraging item multimodal features to model users' interests. However, obtaining multimodal representations for items relies on complex pre-trained encoders, which incurs unacceptable computation cost to train jointly with downstream ranking models. Therefore, it is important to maintain alignment between semantic and behavior space in a lightweight way.

To address these challenges, we propose a Semantic-Behavior Alignment for Cold-start Recommendation framework, which mainly focuses on utilizing multimodal representations that align with the user behavior space to predict CTR. First, we leverage domain-specific knowledge to train a multimodal encoder to generate behavior-aware semantic representations. Second, we use residual quantized semantic ID to dynamically bridge the gap between multimodal representations and the ranking model, facilitating the continuous semantic-behavior alignment. We conduct our offline and online experiments on the Taobao, one of the world's largest e-commerce platforms, and have achieved an increase of 0.83% in offline AUC, 13.21% clicks increase and 13.44% orders increase in the online A/B test, emphasizing the efficacy of our method.

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

Conference acronym 'XX, Woodstock, NY

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-XXXX-X/2018/06 https://doi.org/XXXXXXXXXXXXXXX Yuning Jiang[†] mengzhu.jyn@alibaba-inc.com Alibaba Group Hangzhou, China

CCS Concepts

Information systems → Recommender systems.

Keywords

Recommendation System, Multi-Modal, Click-Through Rate Prediction

ACM Reference Format:

1 Introduction

In recommendation systems, predicting Click-Through Rate (CTR) is of great importance to accurately match users with items that align with their interests. Current recommendation systems [3, 5, 30, 41] primarily rely on features extracted from user-item interaction history to model candidate items and predict CTR, specifically, item ID embeddings, along with statistical features such as historical impression. However, for cold-start and long-tail items, CTR prediction based on these features still faces limitations: ID embeddings are challenging to learn sufficiently because of the lack of training samples, while statistical features are often too sparse, because these items receive few impressions or clicks, as shown in Fig. 1. Consequently, traditional recommendation systems lack the ability to effectively model less popular items, which potentially leads to constrained sales growth and reduced user experience.

To address the limitation in CTR prediction within cold-start and long-tail scenarios, recent works[4, 14, 19, 34] mainly focus on integrating multimodal features to represent both candidate item and users interacted item sequence, further enabling semantic modeling of user interests. With the advancement in multimodal representation learning, various pre-trained multimodal foundation models such as VisualBERT[15], CLIP[22], LLaMA[27], and ViT[8]

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

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

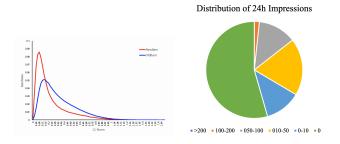


Figure 1: Item ID and statistical features distribution of coldstart items in Taobao.

can generate representation embeddings based on modal information including images, text, and audio. These representations can be further utilized by downstream tasks. In recommendation systems, the multimodal features of items are decoupled from user behavior, with little distinction between popular and cold-start items. This allows recommendation systems to overcome the limitations associated with reliance on behavioral information and to achieve better recommendations for cold-start items[42].

In recent studies the focus is mostly on two parts: (1) how to extract multimodal features aligning with specific recommendation scenario, and (2) how to integrate multimodal features into recommendation system. For extracting multimodal features, most studies[16, 19, 24] finetune pre-trained multimodal models by contrastive learning. This approach encourages that items close in user behavioral space also have similar multimodal representations, which helps bridge the gap between general pre-trained models and specific recommendation tasks. For integrating multimodal features into recommendation systems, some methods[24, 34] capture the semantic user interests from the sequence of items that users have interacted with, then compute the similarity between the interest representation and the multimodal representation of candidate items, further infer whether the user is interested in the candidate item. Other studies employ residual quantization (RQ) method to extract a discretizd, multi-level category-like semantic ID from multimodal feature. [7, 9, 14, 18, 23, 32] use generative models widely used in large language models (LLM) to directly predict the semantic ID of the next item that user will likely interact with based on their historical interactions, while [17, 25, 40] map semantic ID to a semantic embedding to replace randomly hashed item ID, reducing the instability caused by irrelevant items sharing the same item embedding.

Although existing methods have made significant progress in leveraging multimodal features to assist cold-start recommendation, there still remain several challenges and limitations.

Disjoint between trainable ranking model and fixed multimodal embedding.

In industrial-scale recommendation systems, where real-time user requests demand timely updates of parameters and feature embeddings, existing methods[24, 34] face a critical limitation: they follow a 2-stage schema that first extracts item multimodal embeddings through a large pre-trained encoder, and then feeds these

fixed embeddings to downstream ranking models. The complexity of multimodal encoder prevents it from updating in real time, resulting in an increasing gap between multimodal features and dynamically evolving user behavior patterns.

Underutilization of multimodal information.

While some efforts[25, 40] aim to enable trainable multimodal representations with residual quantization (RQ) semantic ID, these methods neglect raw multimodal embeddings, which leads to information loss. Some studies[1, 24] measure modal similarity between a candidate item and the user's interaction sequence. However, the lack of interaction between multimodal feature and other behavioral signals limits the full potential of multimodal representations for user interest modeling.

To tackle these challenges, we propose Semantic-Behavior Alignment for Cold-start Recommendation (SaviorRec) framework, to integrate multimodal information into recommendation model. Specifically, our SaviorRec consists of 3 parts. Firstly, SaviorEnc follows a 2-stage paradigm, deriving behavior-aware semantic representations. Secondly, we design a trainable Modal-Behavior Alignment (MBA) block that ensures consistency between the semantic space and the behavior space, while also preserving the important raw multimodal feature. Thirdly, to integrate multimodal features into the ranking model, we propose a bi-directional target attention mechanism between behavior feature and multimodal feature. This design emphasizes the mutual influence between users' interaction and semantic information, and is beneficial for accurate user interest modeling. We conduct our offline and online experiments on Taobao, one of the world's largest e-commerce platforms, and experiment results show that our SaviorRec effectively incorporates the multimodal information into the ranking model and enhances performance in CTR prediction.

In summary, our main contributions are as follows:

- We train a behavior-aware multimodal encoder to obtain multimodal embeddings and semantic IDs.
- We propose a plug-in module to achieve continuous alignment between semantic and behavior space during the update of ranking model, and further design a bi-directional target attention mechanism to enhance the interaction between behavior signal and semantic information.
- We conduct offline experiments and ablation study to demonstrate the effectiveness of SaviorRec and further validate its performance through online A/B test, where it achieves significant gains.

2 Related Works

2.1 Multimodal Representation Learning

Recent advances in large-scale pre-training have significantly advanced the ability of models to understand multimodal information. These pre-training paradigms can be broadly categorized into two families: reconstructive/generative and contrastive. The generative paradigm trains models to reconstruct masked or missing portions of the input data, thus learning rich, contextualized features. This includes objectives like Masked Language/Image Modeling (MLM/MIM)[2, 11, 31], where models predict hidden text tokens or image patches based on the surrounding multimodal context. Other approaches frame the task as image captioning or text-to-image

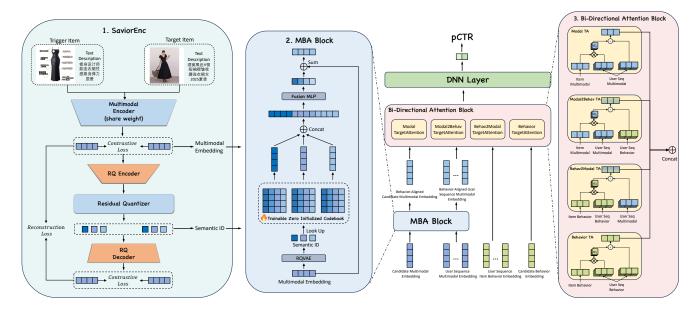


Figure 2: The overview of our SaviorRec framework. SaviorRec mainly consists of three parts. (1) SaviorEnc, (2) MBA block, (3) Bi-directional attention block.

generation [13, 33, 36], forcing the model to understand the deep semantic connections between modalities. While these methods excel at learning an item's objective attributes, their application in recommendation faces a challenge: a model's ability to reconstruct content does not inherently capture user preference. The contrastive paradigm learns a shared embedding space by aligning positive pairs and pushing apart negative ones. This approach [10, 13, 22, 26, 35, 38] has proven highly effective in learning powerful representations from large-scale image-text data. However, a critical gap remains: the general-world alignment learned from Web data often fails to capture the subjective interests of users in e-commerce domain. To address this, our work explicitly uses user co-interaction patterns as the supervisory signal for contrastive tuning, directly embedding behavioral relationships into the semantic space.

2.2 Multimodal Recommendation

Introducing multimodal information into recommendation systems enables comprehensive modeling of items, especially in cold-start and long-tail scenarios. Traditional methods[21] simply use image pixel-level features to enhance representation power. However, recommendation models often lack the ability to fully capture these information. Recent studies are based on a two-stage paradigm: a pre-trained large model is finetuned to generate a multimodal embedding vector for each item, which is then consumed by the downstream recommendation model. SimTier[24] computes a histogram of the cosine similarity between the candidate item and the user behavior sequence as a multimodal feature. QARM[19] uses the quantitative code mechanism to compress multimodal embeddings as semantic IDs for end-to-end training. Though effective, these methods ignore important information contained in original multimodal embeddings. [17, 25, 40] use multimodal semantic ID

to replace item ID, avoiding random collision when hashing item to ID embedding. Other methods[7, 14, 23] quantize multimodal features into discrete IDs, and then use a generative encoder-decoder structure to predict the next item.

3 Methodology

3.1 Task Definition and Overview

3.1.1 Task Definition for Cold-Start CTR Prediction. The Click-Through Rate (CTR) prediction task is to predict the probability of a user clicking a candidate item using a ranking model. Specifically, in the cold-start recommendation scenario of Taobao, for each user $u \in \mathcal{U}$, we use a unique user ID ID_u , user profile P_u and items that the user interacted with $Seq_u = [item_u^1, ..., item_u^n]$ to model the user's interest. And for each candidate item $i \in I$ we construct a series of features including item ID ID_i , statistical features extracted from user interactions S_i and multimodal feature MM_i . In addition, we leverage a base CTR score $pCTR_{u,i}^{base}$ predicted by the model of main recommendation scenario, which, although not very accurate for cold-start and long-tail items, still serves as a valuable reference.

Based on the input features above, the cold-start CTR prediction task can be formulated as:

$$pCTR_{u,i} = f([ID_u, P_u, Seq_u], [ID_i, S_i, MM_i], pCTR_{u,i}^{base})$$
(1)

3.1.2 Method Overview. The overall structure of our method is shown in Fig.2, which takes features described in 3.1.1 as inputs and predicts CTR score through a deep neural network. Our method consists of three important parts. SaviorEnc extracts multimodal embedding and residual quantized semantic ID from image and text description of items. Modal-behavior alignment module (MBA) utilizes semantic ID to adjust the multimodal embedding, ensuring the alignment between semantic and behavior space. Bi-directional target attention block extracts user interests from the behavior

sequence, and achieves information fusion between behavioral and multimodal features.

3.2 SaviorEnc

A key challenge in multimodal recommendation is the semantic gap between the pre-training objectives and the user interests in downstream tasks. To address this, we design SaviorEnc, a multimodal encoder following a 2-stage schema that learns multimodal representations directly from user-interaction signals. Stage 1 trains a powerful encoder to extract multimodal embedding and stage 2 employsa RQ-VAE to generate corresponding semantic ID.

To obtain powerful multimodal encoders, we initialize our model with the open-source CN-CLIP [35] and perform a domain adaptation fine-tuning, training CN-CLIP on our proprietary product dataset. This serves as the foundation for the behavior-aware encoder.

At stage 1, we construct a set of positive item pairs \mathcal{P}^+ by mining frequent co-click patterns (i,j) from user logs. This data-driven approach allows us to capture both intrinsic item similarity and complementarity. Further, we train a contrastive model with the co-click pairs. For each item i, we obtain unimodal representations from a vision encoder $f_v(v_i)$ and a text encoder $f_t(t_i)$, which are then fed into a multimodal transformer to capture cross-modal interactions, yielding a fused representation. The output is subsequently mapped to a latent space via an MLP projection head, $g_{proj}(\cdot)$, for the contrastive loss calculation:

$$\mathbf{z}_i = g_{proj}(g([f_v(v_i); f_t(t_i)])) \in \mathbb{R}^d$$
 (2)

The model is optimized using an InfoNCE loss [28], which maximizes the agreement between positive pairs (i, j) against in-batch negatives. The loss for the item i can be formulated as:

$$\mathcal{L}_{i} = -\log \frac{\exp(\operatorname{sim}(\mathbf{z}_{i}, \mathbf{z}_{j})/\tau)}{\exp(\operatorname{sim}(\mathbf{z}_{i}, \mathbf{z}_{j})/\tau) + \sum_{k \neq i, j} \exp(\operatorname{sim}(\mathbf{z}_{i}, \mathbf{z}_{k})/\tau)} \quad (3)$$

where $sim(\cdot, \cdot)$ is cosine similarity and τ is temperature. Through these training objectives, we obtain a dense multimodal feature \mathbf{z}_i for each item. The resulting embedding space is semantically aligned with user behaviors, where co-interacted items are mapped closer together.

While the multimodal representation z_i is effective, it still presents a challenge for ranking models, as these large, frozen embeddings cannot be fine-tuned. To enable dynamic updating and continuous behavior alignment of the multimodal embeddings detailed in 3.3, at stage 2 we discretize z_i into a sequence of semantic IDs, inspired by [6, 12].

Specifically, we employ a residual quantized variational autoencoder (RQ-VAE) [37] which learns to map a continuous vector to a sequence of discrete codes $\mathbf{c}=(c_i^1,c_i^2,...,c_i^L)$ from L different codebooks $C=\{C_1,...,C_L\}$. The quantization is performed residually. The first quantizer finds the closest code for the input vector, and the second quantizer finds the code for the residual result of the first. This process continues iteratively for subsequent quantizers.

$$\mathbf{r}_{i}^{1} = \operatorname{Enc}(\mathbf{z}_{i})$$

$$c_{i}^{l} = \operatorname{argmin}_{c} \|\mathbf{r}_{i}^{l} - \mathbf{C}_{l}^{c}\|_{2} \quad \text{for } l = 1, ..., L$$

$$\mathbf{r}_{i}^{l+1} = \mathbf{r}_{i}^{l} - \mathbf{C}_{i}^{c_{i}^{l}}$$

$$(4)$$

The sequence of indices of the vectors $\{\mathbf{C}_l^{c_l^i}\}_{l=1}^L$ forms the semantic ID for item i. The reconstructed vector is $\hat{\mathbf{z}}_i = \mathrm{Dec}(\sum_{l=1}^L \mathbf{C}_l^{c_l^i})$.

To improve codebook utilization and prevent collapse, we replace the argmin assignment with an optimal transport-based Sinkhorn algorithm [39], which encourages a more uniform distribution of code usage. Furthermore, to ensure that the generated IDs retain semantic coherence, we introduce an additional contrastive objective. Using the same positive pairs \mathcal{P}^+ , we apply an auxiliary InfoNCE loss to the reconstructed vectors $\hat{\mathbf{z}}_i$ and $\hat{\mathbf{z}}_j$. The final training objective for the RQ-VAE is:

$$\mathcal{L}_{\text{RQ-VAE}} = \lambda_0 \mathcal{L}_{\text{reconstruct}} + \lambda_1 \mathcal{L}_{\text{commit}} + \mathcal{L}_{\text{contrast}}(\hat{\mathbf{z}}_i, \hat{\mathbf{z}}_j) \quad (5)$$

Following this paradigm, SaviorEnc yields a behavior-aware multimodal embedding and a residual semantic ID for each item, which together serve as input features for the ranking model.

3.3 Modal-Behavior Alignment Module

After obtaining the preliminary behavior-aligned item multimodal embedding, it is crucial to design an effective way to integrate it into the ranking model. However, since the item multimodal features extracted by the SaviorEnc are computationally expensive to update in real time, the discrepancy between the behavior space of the ranking model and the semantic space of the multimodal embeddings tends to grow as the ranking model is trained and updated. Therefore, we propose a Modal-Behavior Alignment (MBA) module to enhance the alignment between the semantic and behavior space.

For each item i, MBA module takes raw item multimodal embedding \mathbf{z} and RQ code $\mathbf{c} = [c_1, ..., c_L]$ mentioned in 3.2 as input (subscript i is omitted for simplicity). Residual quantization (RQ) is a method compressing embedding vectors into hierarchical cluster IDs, where each layer of IDs captures semantic information coarse to fine. Based on the RQ code, MBA module constructs a behavioralignment vector, which serves as a complementary signal to bridge the gap between the frozen multimodal embedding and the dynamically updated ranking model. In practice, we firstly construct a trainable zero-initialized MBA codebook $C = [C_1, ..., C_L]$ with the same shape as the RQ codebook obtained in 3.2, and each RQ code $c_l \in \mathbf{c}$ corresponds to a embedding vector \mathbf{v}_l at the l_{th} layer of MBA codebook.

$$\mathbf{v}_l = C_l^{c_l} \tag{6}$$

$$\mathbf{v} = [\mathbf{v}_1, ..., \mathbf{v}_L]$$
 (7)

After we gather the embedding vectors $\mathbf{v} = [\mathbf{v}_1, ..., \mathbf{v}_L]$ from MBA codebook, we propose a way different from traditional RQ to reconstruct multimodal embedding. Following the practice[6, 12] in image and audio domains, recent researches [17, 40] directly sum the embedding vectors in \mathbf{v} to produce the multimodal feature. However, this approach can lead to instability in codebook training, since the gradients back-propagated to \mathbf{v}_l for each $l \in \{1, ..., L\}$ are

identical, whereas the semantic codes follow a coarse-to-fine hierarchy and thus have different levels of importance and numeric scales. Therefore, we propose a fusion layer to adaptively learn the importance of embeddings from each residual layer, further enabling soft optimization of the codebook during gradient back-propagation. In practice, we first concatenate each embedding vector, then apply L2 normalization and use an MLP layer to fuse embeddings from each residual layer and map the result back to the original dimension.

$$\mathbf{v}_{align} = MLP(Concat([\mathbf{v}_1, ..., \mathbf{v}_L]))$$
(8)

The output embedding vector of the MLP layer is further added to the original multimodal embedding **z**, which serves as a residual signal to align multimodal embedding to behavior space.

$$\mathbf{z}_{aliqn} = \mathbf{z} + \mathbf{v}_{aliqn} \tag{9}$$

By combining the codebook-based alignment signal with the original embedding in a skip-connect pattern, it is possible to avoid the loss of multimodal information extracted in 3.2, while also helping the codebook focus on the most important content. The final output $\mathbf{z}_{\text{align}}$ of MBA module will serve as behavior-aligned multimodal embedding of candidate item or items in user behavior sequence and be fed into the bi-directional target attention module.

With the help of the trainable MBA codebook, the multimodal embedding can be updated along with the training of the whole ranking model, overcoming the restrictions caused by the semanticbehavior gap.

3.4 Bi-Directional Target Attention Mechanism

For the behavior feature embeddings including item IDs and statistical features $\mathbf{h}_{seq} = [\mathbf{h}_1, ..., \mathbf{h}_n] = Concat([\mathbf{ID}_1, ..., \mathbf{ID}_n], [\mathbf{S}_1, ..., \mathbf{S}_n])$ and multimodal embeddings $\mathbf{z}_{seq} = [\mathbf{z}_1^{align}, ..., \mathbf{z}_1^{align}]$ of user interaction sequence, we propose a bi-directional target attention mechanism to fuse information from both the modal-side and the behavior-side, further to extract user interests. This promotes interaction between the behavior embedding and multimodal embedding, thus enhancing the representation power of the model.

Target attention(TA) mechanism[41] calculates a similarity score between user sequence and candidate item. Then the embeddings in the behavior sequence are aggregated based on similarity scores to extract the user's interest.

$$TA(Q, K, V) = \left(\frac{(\mathbf{Q}\mathbf{W}^q)(\mathbf{K}\mathbf{W}^k)^T}{\sqrt{d}}\right)(\mathbf{V}\mathbf{W}^v)$$
 (10)

Our bi-directional target attention mechanism consists of four TA blocks, including two regular TA blocks which are applied separately to the behavior embedding and the multimodal embedding, in order to extract user interest in the behavior space and the semantic space. This process can be formulated as follows, where subscript *cand* represents the candidate item, and *m* and *b* refer to modal and behavior respectively.

$$\mathbf{h}_b = TA(\mathbf{h}_{cand}, \mathbf{h}_{seq}, \mathbf{h}_{seq}) \tag{11}$$

$$\mathbf{h}_{m} = TA(\mathbf{z}_{cand}, \mathbf{z}_{seq}, \mathbf{z}_{seq}) \tag{12}$$

The other two TA blocks aim to fuse the modal and behavior embeddings and capture cross-information. At Modal2Behavior TA block, we use the similarity score in semantic space to aggregate

Table 1: The statistic of evaluation dataset (%). PV refers to Pages View of a candidate item.

PV Group	Samples	Clicks	Items
[0, 100)	2.24	2.16	31.07
[100, 500)	17.09	17.74	33.98
[500, 1000)	32.29	31.01	24.27
[1000, 5000)	24.90	22.39	8.74
[5000, 10000)	8.66	8.79	0.87
[10000, 20000)	14.15	16.75	0.68
$[20000, \infty)$	0.67	1.16	0.39

behavior embedding sequence. Similarly, we use the similarity score in behavior space to aggregate multimodal embedding sequence at Behavior2Modal TA block.

$$\mathbf{h}_{m2b} = TA(\mathbf{z}_{cand}, \mathbf{z}_{sea}, \mathbf{h}_{sea}) \tag{13}$$

$$\mathbf{h}_{b2m} = TA(\mathbf{h}_{cand}, \mathbf{h}_{seq}, \mathbf{z}_{seq}) \tag{14}$$

Finally, we concatenate the outputs of the TA blocks to form the feature vector for the CTR prediction network.

$$\mathbf{h}_{BiD-TA} = Concat([\mathbf{h}_b, \mathbf{h}_m, \mathbf{h}_{b2m}, \mathbf{h}_{m2b}]) \tag{15}$$

$$pCTR = DNN(\mathbf{h}_{BiD-TA}, other_features)$$
 (16)

We optimize the CTR prediction model, including the MBA module and bi-directional attention mechanism, by cross-entropy loss:

$$Loss = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(pCTR_i) + (1 - y_i) \log(1 - pCTR_i)]$$
 (17)

where N is batch size, $y_i \in \{0, 1\}$ is the label of sample i, and pCTR is the output of CTR prediction model.

4 Experiment

To assess the effectiveness of SaviorRec, we conduct experiments around the following key research questions.

- RQ1: How does our model perform compared to other methods that leverage multimodal information?
- **RQ2**: What is the impact of our model's different components on overall performance, and how does it address the limitations of existing methods?
- **RQ3**: How can we achieve a trade-off between the parameter size and effectiveness of multimodal modeling?
- RQ4: What roles do behavioral and semantic information play in user interest modeling and item recommendation?

4.1 Experimental Settings

4.1.1 Metrics. (1)CTR Prediction: For CTR prediction task, we use AUC as the main evaluation metric. AUC is the probability that a positive user-item pair receives a higher score than a negative user-item pair, which reflects the model's scoring ability. For a more fine-grained evaluation across items with varying popularity, we partition items according to their total historical page views (PV) and compute the AUC within each group. (2)Multimodal Representation Learning: In addition, we directly employ a behavioral retrieval task that simulates an item-to-item (i2i) recommendation

Methods	Total AUC	AUC across item PV Buckets						
Wicthous		[0,100)	[100,500)	[500,1000)	[1000,5000)	[5000,10000)	[10000,20000)	[20000,∞)
Base	71.28	70.34	70.16	70.67	71.12	73.47	72.01	71.93
BBQRec	71.61	71.08	70.65	71.05	71.41	73.62	72.16	71.93
CHIME	71.21	70.27	70.07	70.60	71.06	73.41	71.97	71.87
MIM	72.02	71.71	71.20	71.50	71.82	73.92	72.48	72.02
SimTier	71.36	70.28	70.23	70.76	71.22	73.52	72.03	71.79
SaviorRec	72.11	71.87	71.32	71.61	71.89	73.95	72.50	72.04

Table 2: The AUC(%, ↑) results of SaviorRec and baselines. The best is denoted in bold font.

Table 3: Ablation study of different SaviorRec components, where "w/o" is short for "without". The best is denoted in bold font.

Methods	Total AUC	Δ
Base	71.28	-0.83
w/o MBA	72.00	-0.11
w/o multimodal embedding	71.80	-0.31
w/o Bi-Dirc Attn	71.98	-0.13
SaviorRec	72.11	-

scenario to evaluate our multimodal encoder. For each historical item in a user's sequence, we retrieve the top-K most similar items from the corpus using its feature vector. The ground truth consists of the items the user subsequently interacted with. We use the following standard metrics for this evaluation: Hitrate@K, which measures the proportion of ground-truth items that appear in the top-K retrieved list.

- 4.1.2 Baselines. We select Taobao's cold-start model currently used online, and several studies on the application of multimodal features in recommendation systems as baselines. Considering there are significant differences among some baseline methods in the overall paradigm, we only use the corresponding modules to construct multimodal features.
 - Base model is Taobao's current cold-start CTR prediction model not utilizing multimodal features; it only consists of target attention over the Item ID and other statistical features.
 - **BBQRec**[14] proposes an auxiliary module within the selfattention through a non-invasive manner.
 - **CHIME**[1] designs an interest compression module that endto end learns the user interest distribution and compresses it as a compact histogram.
 - MIM[34] proposes a fusion interest module to combine item ID and content interests.
 - **SimTier**[24] computes the similarities between the candidate item and the user's interacted items, and summarizes them in a histogram to serve as a multimodal feature.
- 4.1.3 Datasets. Experiments are conducted on traffic logs collected from Taobao's homepage feed. We construct an industrial-scale

cold-start dataset for training and evaluation according to Taobao's cold-start item selection algorithm, which is detailed in Tab.1. We choose data from a three-week period in July 2025 for training, with the final day for evaluation, which contains on the order of 10^8 samples each day.

4.1.4 Implement Details. For the multimodal representation learning, we use the CN-CLIP [35] architecture as our base model. We then employ a 3-layer multimodal transformer for fusing the single-model outputs. This feature generation model is trained for 10 epochs with a learning rate of 1e-4, a 1-epoch warmup period, and a batch size of 4096.

For the **semantic ID training stage**, we train the RQ-VAE module. The training runs for 150 epochs with a learning rate of 0.002 and a large batch size of 16384 to ensure stable codebook learning. The loss weights for the reconstruction objectives and commitment loss are set to $\lambda_0 = 1000$ and $\lambda_1 = 0.5$, respectively. We set the RQ codebook size of each layer K = 2048, codebook length L = 8, and embedding dimension d = 64.

For **the CTR prediction model**, we set the batch size to 1024 and use AdagradV2 as the optimizer, with a learning rate that gradually decays from 0.01 to 0.001. The MBA codebook and the RQ codebook are of the same size. In the bi-directional target attention module, we use a multi-head attention mechanism [29] with 8 heads. The final DNN part contains 3 MLP layers to predict CTR and uses LeakyReLU [20] as the activation function.

4.2 Overall Performance (RQ1)

In this section, we compared our method with baseline recommendation models using multimodal features. The overall performance is shown in Tab.2 and our method outperforms other baselines at both total AUC and AUC across item PV Buckets.

Compared to the base model without modality information, most methods achieve significant AUC improvement, demonstrating the substantial advantages of semantic information for user interest modeling and item recommendation. However CHIME fails to outperform the base model because its way of constructing multimodal feature embedding does not align well with the ranking model, which prevents it from capturing useful information.

Among the multimodal recommendation methods, SaviorRec performs best in all metrics, which highlights the effectiveness of the MBA module and bi-directional target attention mechanism.

Table 4: Ablation study on multimodal representation learning.

Model	Domain Adapt.	i2i Alignment	Hit@30 (%)
#1 (Official)			28.56
#2	✓		32.36
#3		\checkmark	39.28
#4 (Ours)	\checkmark	\checkmark	41.30

Compared to methods extracting statistical features such as histograms (SimTier, CHIME), methods that directly use attention mechanisms (SaviorRec, MIM) achieve better performance. This shows that attention can preserve more multimodal information and offers clear advantages.

For AUC across different item historical page views(PVs) groups, the lower the PV, the better our method performs. This indicates that multimodal features play a more important role in the recommendation of cold-start items, due to the lack of behavioral information and as a result, insufficient modeling of item IDs.

4.3 Ablation and Analysis (RQ2)

We conduct ablation studies to validate the effectiveness of each component in SaviorRec as follows:

- Base: Only consists of target attention over the Item ID and subsequent MLP layers.
- w/o MBA: We remove the MBA module and directly feed the frozen multimodal embeddings into bi-directional attention block
- w/o multimodal embedding: We initialize the codebook in MBA module with that obtained in the RQ stage, and do not sum the codebook embedding with the original multimodal embedding.
- w/o Bi-Dirc Attn: We remove the bi-directional attention block and only apply target attention over multimodal features

The experiment results are shown in Tab.3. (1) Our MBA module can add a dynamic alignment signal to the static multimodal embedding, resulting in a 0.11% improvement in AUC. However, when we initialize the codebook in the MBA module with that obtained in the RQ stage, instead of zero-initializing it and summing the codebook embedding with the original multimodal embedding, the AUC decreases by 0.31%. This demonstrates that the skip-connect structure we designed in the MBA module helps maintain a balance between the multimodal features and the modal-behavior alignment signal, preventing the behavioral signal from overwhelming the original multimodal information during training. (2) As we remove the bi-directional attention block, total AUC decreases by 0.13%, indicating that the mutual fusion between semantic and behavioral signal is beneficial for user-item modeling.

To verify that our zero-initialized MBA codebook can effectively learn alignment signal, we analyze the average L2 norm of codebook embeddings and subsequent fusion MLP weights of each residual layer in Fig.3. The MBA codebook is capable of learning the relative importance across different layers, exhibiting a coarse-to-fine feature hierarchy. In the fusion MLP layer, the network weights

corresponding to each residual layer also show a similar importance distribution. This ensures that, during training, the gradients of each codebook layer are scaled in accordance with their importance, thereby avoiding the training instability caused by direct sum pooling of the codebook.

Additionally, we also conduct an ablation study on SaviorEnc's key components. We evaluate the quality of the representations using a behavioral retrieval task that simulates a real-world itemto-item (i2i) recommendation scenario. For each user, we take their historical clicked items as queries. We then leverage the corresponding multimodal representations to retrieve the top-30 most similar items from the entire item corpus. A retrieval is considered successful (as a "hit") if the retrieved list contains the item the user clicked after the query item. The results of our ablation study are presented in Table 4. The results demonstrate the effectiveness of our approach. Domain adaptation (#2 vs. #1) yields a notable improvement, while i2i behavior alignment (#3 vs. #1) proves more critical, delivering a 10.72% absolute gain in Hit@30. Our full SaviorEnc (#4) achieves the best performance by combining these two complementary strategies, validating the overall design.

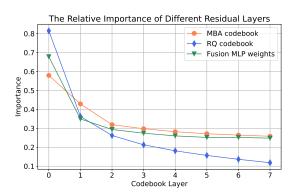


Figure 3: The relative importance of different residual layers. The importance scores are obtained by normalizing the average L2 norm of the corresponding parameters.

4.4 Parameter Analysis of MBA Module Codebook (RQ3)

In this section, we reduce the embedding dimension of the MBA codebook to investigate whether a balance can be achieved between model performance and the number of codebook parameters. The result is shown in Tab.5. In our original MBA Module, the embedding dimension is set to 64, consistent with the codebook used in the RQ stage. We first reduce the embedding dimension to 32 and 16, and the AUC only decreases slightly. However, when the dimension is further reduced to 8, the AUC drops more noticeably, approaching the performance of the model which only uses frozen multimodal embedding without the MBA module, but still better than baselines in 4.2. These results indicate that our model can be made more lightweight with minimal performance loss, which is valuable for deployment in industrial settings. However, reducing the codebook dimension too much will constrain the modeling capability.

Table 5: AUC(%) with different codebook dimensions

MBA Codebook Dimension	64	32	16	8
Total AUC	72.11	72.07	72.08	72.03

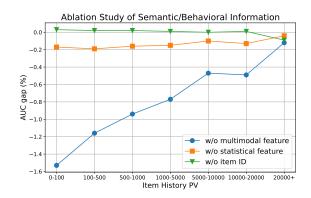


Figure 4: The AUC drop after removing ID, statistical and multimodal features.

4.5 Effectiveness of Behavioral and Semantic Information (RQ4)

In SaviorRec, we use ID embedding, statistical features and multimodal features to model both candidate item and items in user behavior sequence. As shown in Fig.4, we separately remove these features to explore the role of behavioral and semantic information in recommendation.

When we remove all multimodal features and the corresponding modules in our method, the AUC drops significantly across all PV intervals, demonstrating the importance of semantic information for cold-start scenario. For items with lower PV, the lack of user interactions leads to sparse statistical features and insufficient training of item IDs, therefore removing multimodal features has a greater negative impact.

For features in the behavior space, removing the statistical features from the bi-directional attention module leads to a smaller drop in AUC, and are equally important for both popular and cold-start items. However, item ID contributes little to cold-start recommendation and even has a negative impact when PV < 5000, indicating that when the item PV is low, training the item ID embedding with a small number of samples will only lead to unstable representations. When item PV > 20000, the drop in AUC without item ID suggests that it is able to effectively capture behavioral information for these popular items.

Fig.5 visualizes the multimodal embedding spaces using t-SNE, with several *Harry Potter* themed items (books, robes, scarves, wands) highlighted. The baseline "Official CLIP" model fails to group items with similar use interaction pattern. Our "SaviorRec", trained with user behavior signals, forms a single cluster for all *Harry Potter* items. This demonstrates that the behavior alignment mechanism in SaviorEnc and MBA module can capture user interests.



Figure 5: Visualization of item multimodal embedding space. SaviorRec achieves alignment among items of different categories under the same behavioral paradigm.

Table 6: Results of online A/B test.

-	Metrics	Clicks	Orders	CTR
_	Impr.(%)	13.31	13.44	12.80

4.6 Online A/B Test

We deployed SaviorRec and conducted A/B test on Taobao's "Guess You Like" service, specifically targeting cold-start items identified by predefined rules. Clicks, orders, and CTR are adopted as the main evaluation metrics, since they are the primary indicators of interest in our industrial setting. Tab.6 shows the online results, where our method achieved a 13.31% increase in clicks, a 13.44% increase in orders, and a 12.80% improvement in CTR. This highlights the importance of our method in modeling cold-start items and enabling accurate recommendations.

5 Conclusion

In this paper, we propose SaviorRec, a novel and deployable multimodal recommendation framework that achieves alignment between semantic information and user behavior. SaviorEnc leverages co-click item pairs to generate behavior-aware multimodal representation and semantic ID. The MBA module achieves continuous alignment between semantic information and user behavior throughout the ranking model's training by dynamically updating a residual codebook. The bi-directional target attention mechanism promotes the fusion of behavior and multimodal features, and enhances the model's representation power. We conducted extensive experiments to demonstrate that SaviorRec outperforms existing methods for multimodal recommendation, verifying the effectiveness of our method. The substantial gains in online clicks, orders, and CTR demonstrate that our model is capable of utilizing behavior-aligned multimodal information for cold-start item modeling and high-quality recommendation in real-world industrial settings.

References

- Yong Bai, Rui Xiang, Kaiyuan Li, Yongxiang Tang, Yanhua Cheng, Xialong Liu, Peng Jiang, and Kun Gai. 2025. CHIME: A Compressive Framework for Holistic Interest Modeling. arXiv:2504.06780 [cs.IR] https://arxiv.org/abs/2504.06780
- [2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEiT: BERT Pre-Training of Image Transformers. arXiv:2106.08254 [cs.CV] https://arxiv.org/abs/ 2106.08254
- [3] Jianxin Chang, Chenbin Zhang, Zhiyi Fu, Xiaoxue Zang, Lin Guan, Jing Lu, Yiqun Hui, Dewei Leng, Yanan Niu, Yang Song, and Kun Gai. 2023. TWIN: TWo-stage Interest Network for Lifelong User Behavior Modeling in CTR Prediction at Kuaishou. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Long Beach, CA, USA) (KDD '23). Association for Computing Machinery, New York, NY, USA, 3785–3794. doi:10.1145/3580305.3599922
- [4] Gaode Chen, Ruina Sun, Yuezihan Jiang, Jiangxia Cao, Qi Zhang, Jingjian Lin, Han Li, Kun Gai, and Xinghua Zhang. 2024. A Multi-modal Modeling Framework for Cold-start Short-video Recommendation. In Proceedings of the 18th ACM Conference on Recommender Systems (Bari, Italy) (RecSys '24). Association for Computing Machinery, New York, NY, USA, 391–400. doi:10.1145/3640457.3688098
- [5] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In Proceedings of the 10th ACM Conference on Recommender Systems (Boston, Massachusetts, USA) (RecSys '16). Association for Computing Machinery, New York, NY, USA, 191–198. doi:10.1145/2959100. 2959190
- [6] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=ivCd8z8zR2 Featured Certification, Reproducibility Certification.
- [7] Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. OneRes: Unifying Retrieve and Rank with Generative Recommender and Iterative Preference Alignment. arXiv:2502.18965 [cs.IR] https://arxiv.org/abs/2502.18965
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV] https://arxiv.org/abs/2010.11929
- [9] Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2023. Learning Vector-Quantized Item Representation for Transferable Sequential Recommenders. arXiv:2210.12316 [cs.IR] https://arxiv.org/abs/2210.12316
- [10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. arXiv:2102.05918 [cs.CV] https://arxiv.org/abs/2102.05918
- [11] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Visionand-Language Transformer Without Convolution or Region Supervision. arXiv:2102.03334 [stat.ML] https://arxiv.org/abs/2102.03334
- [12] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive Image Generation using Residual Quantization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, 11513-11522. doi:10.1109/CVPR52688.2022.01123
- [13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv:2201.12086 [cs.CV] https://arxiv.org/abs/2201.12086
- [14] Kaiyuan Li, Rui Xiang, Yong Bai, Yongxiang Tang, Yanhua Cheng, Xialong Liu, Peng Jiang, and Kun Gai. 2025. BBQRec: Behavior-Bind Quantization for Multi-Modal Sequential Recommendation. arXiv:2504.06636 [cs.IR] https://arxiv.org/ abs/2504.06636
- [15] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557 [cs.CV] https://arxiv.org/abs/1908.03557
- [16] Youhua Li, Hanwen Du, Yongxin Ni, Pengpeng Zhao, Qi Guo, Fajie Yuan, and Xiaofang Zhou. 2024. Multi-Modality is All You Need for Transferable Recommender Systems. In 2024 IEEE 40th International Conference on Data Engineering (ICDE). IEEE Computer Society, Los Alamitos, CA, USA, 5008–5021. doi:10.1109/ICDE60146.2024.00380
- [17] Guanyu Lin, Zhigang Hua, Tao Feng, Shuang Yang, Bo Long, and Jiaxuan You. 2025. Unified Semantic and ID Representation Learning for Deep Recommenders. arXiv:2502.16474 [cs.IR] https://arxiv.org/abs/2502.16474
- [18] Han Liu, Yinwei Wei, Xuemeng Song, Weili Guan, Yuan-Fang Li, and Liqiang Nie. 2024. MMGRec: Multimodal Generative Recommendation with Transformer Model. arXiv:2404.16555 [cs.IR] https://arxiv.org/abs/2404.16555
- [19] Xinchen Luo, Jiangxia Cao, Tianyu Sun, Jinkai Yu, Rui Huang, Wei Yuan, Hezheng Lin, Yichen Zheng, Shiyao Wang, Qigen Hu, Changqing Qiu, Jiaqi Zhang, Xu Zhang, Zhiheng Yan, Jingming Zhang, Simin Zhang, Mingxing Wen, Zhaoje Liu, Kun Gai, and Guorui Zhou. 2024. QARM: Quantitative Alignment Multi-Modal Recommendation at Kuaishou. arXiv:2411.11739 [cs.IR] https://arxiv.org/abs/

- 2411 11739
- [20] A.L. Maas, A.Y. Hannun, and A.Y. Ng. 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In Proceedings of the International Conference on Machine Learning. Atlanta, Georgia.
- [21] Kaixiang Mo, Bo Liu, Lei Xiao, Yong Li, and Jie Jiang. 2015. Image feature learning for cold start problem in display advertising. In Proceedings of the 24th International Conference on Artificial Intelligence (Buenos Aires, Argentina) (IJCAI'15). AAAI Press, 3728–3734.
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/ abs/2103.00020
- [23] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. 2023. Recommender Systems with Generative Retrieval. arXiv:2305.05065 [cs.IR] https://arxiv.org/abs/2305.05065
- [24] Xiang-Rong Sheng, Feifan Yang, Litong Gong, Biao Wang, Zhangming Chan, Yujing Zhang, Yueyao Cheng, Yong-Nan Zhu, Tiezheng Ge, Han Zhu, Yuning Jiang, Jian Xu, and Bo Zheng. 2024. Enhancing Taobao Display Advertising with Multimodal Representations: Challenges, Approaches and Insights. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (Boise, ID, USA) (CIKM '24). Association for Computing Machinery, New York, NY, USA, 4858–4865. doi:10.1145/3627673.3680068
- [25] Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, Ed Chi, and Xinyang Yi. 2024. Better Generalization with Semantic IDs: A Case Study in Ranking for Recommendations. In Proceedings of the 18th ACM Conference on Recommender Systems (Bari, Italy) (RecSys '24). Association for Computing Machinery, New York, NY, USA, 1039–1044. doi:10.1145/3640457. 3681100
- [26] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. arXiv:2303.15389 [cs.CV] https://arxiv.org/abs/2303.15389
- [27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] https://arxiv.org/abs/2302.13971
- [28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748 [cs.LG] https://arxiv.org/ abs/1807.03748
- [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA) (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.
- [30] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-scale Learning to Rank Systems. In Proceedings of the Web Conference 2021 (Ljubljana, Slovenia) (WWW '21). Association for Computing Machinery, New York, NY, USA, 1785–1797. doi:10.1145/3442381.3450078
- [31] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. arXiv:2208.10442 [cs.CV] https://arxiv.org/abs/2208. 10442
- [32] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Learnable Item Tokenization for Generative Recommendation. arXiv:2405.07314 [cs.IR] https://arxiv.org/abs/2405.07314
- [33] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. arXiv:2108.10904 [cs.CV] https://arxiv.org/abs/2108.10904
- [34] Bencheng Yan, Si Chen, Shichang Jia, Jianyu Liu, Yueran Liu, Chenghan Fu, Wanxian Guan, Hui Zhao, Xiang Zhang, Kai Zhang, Wenbo Su, Pengjie Wang, Jian Xu, Bo Zheng, and Baolin Liu. 2025. MIM: Multi-modal Content Interest Modeling Paradigm for User Behavior Modeling. arXiv:2502.00321 [cs.IR] https: //arxiv.org/abs/2502.00321
- [35] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2023. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. arXiv:2211.01335 [cs.CV] https://arxiv.org/abs/2211.01335
- [36] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv:2205.01917 [cs.CV] https://arxiv.org/abs/2205.01917
- [37] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. SoundStream: An End-to-End Neural Audio Codec. arXiv:2107.03312 [cs.SD] https://arxiv.org/abs/2107.03312

- [38] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. arXiv:2303.15343 [cs.CV] https://arxiv.org/abs/2303.15343
- [39] Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. 2024. Preventing Local Pitfalls in Vector Quantization via Optimal Transport. arXiv:2412.15195 [cs.CV] https://arxiv.org/abs/2412.15195
- [40] Carolina Zheng, Minhui Huang, Dmitrii Pedchenko, Kaushik Rangadurai, Siyu Wang, Gaby Nahum, Jie Lei, Yang Yang, Tao Liu, Zutian Luo, Xiaohan Wei, Dinesh Ramasamy, Jiyan Yang, Yiping Han, Lin Yang, Hangjun Xu, Rong Jin, and Shuang Yang. 2025. Enhancing Embedding Representation Stability in Recommendation Systems with Semantic ID. arXiv:2504.02137 [cs.IR] https://px.doi.org/10.1016/j.jan.2016.02137
- //arxiv.org/abs/2504.02137
- [41] 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 (London, United Kingdom) (KDD '18). Association for Computing Machinery, New York, NY, USA, 1059–1068. doi:10.1145/3219819.3219823
- [42] Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. 2021. Fairness among New Items in Cold Start Recommender Systems (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 767–776. doi:10. 1145/3404835.3462948