Enhancing Taobao Display Advertising with Multimodal Representations: Challenges, Approaches and Insights

Xiang-Rong Sheng[*]
Feifan Yang[*]
Litong Gong[*]
Biao Wang[*]
xiangrong.sxr@alibaba-inc.com
yangfeifan.yff@alibaba-inc.com
gonglitong.glt@alibaba-inc.com
eric.wb@alibaba-inc.com
Alibaba Group
Beijing, China

Zhangming Chan
Yujing Zhang
Yueyao Cheng
Yong-Nan Zhu
zhangming.czm@alibaba-inc.com
jinghan.zyj@alibaba-inc.com
yueyao.syy@alibaba-inc.com
yongnan.zy@alibaba-inc.com
Alibaba Group
Beijing, China

Tiezheng Ge
Han Zhu[†]
Yuning Jiang
Jian Xu
Bo Zheng
tiezheng.gtz@alibaba-inc.com
zhuhan.zh@alibaba-inc.com
mengzhu.jyn@alibaba-inc.com
xiyu.xj@alibaba-inc.com
bozheng@alibaba-inc.com
Alibaba Group
Beijing, China

ABSTRACT

Despite the recognized potential of multimodal data to improve model accuracy, many large-scale industrial recommendation systems, including Taobao display advertising system, predominantly depend on sparse ID features in their models. In this work, we explore approaches to leverage multimodal data to enhance the recommendation accuracy. We start from identifying the key challenges in adopting multimodal data in a manner that is both effective and cost-efficient for industrial systems. To address these challenges, we introduce a two-phase framework, including: 1) the pre-training of multimodal representations to capture semantic similarity, and 2) the integration of these representations with existing ID-based models. Furthermore, we detail the architecture of our production system, which is designed to facilitate the deployment of multimodal representations. Since the integration of multimodal representations in mid-2023, we have observed significant performance improvements in Taobao display advertising system. We believe that the insights we have gathered will serve as a valuable resource for practitioners seeking to leverage multimodal data in their systems.

CCS CONCEPTS

• **Information systems → Information retrieval**.

KEYWORDS

Multimodal Representations, Recommendation System

---

[*]Equal contribution.
[†]Corresponding author.

1 INTRODUCTION

Traditionally, the recommendation models employed in Taobao's display advertising system, as with many other industrial systems, have largely relied on discrete IDs as features. Despite their widespread use, ID-based models have intrinsic drawbacks, such as the inability to capture the semantic information contained within multimodal data.

To address these issues, certain industrial systems have attempted to incorporate multimodal data into the ID-based models [26, 29]. Typically, these approaches employ a two-phase framework, including 1) acquiring multimodal representations through either generic or scenario-specific pre-training, and 2) integrating these representations into the recommendation models.

Despite these advancements, a significant number of industrial systems still depend exclusively on ID features. This is often attributed to the concern that the performance gains from multimodal data might not compensate for the costs involved in their deployment. These costs encompass pre-training multimodal encoders, incrementally generating representations for new items, and other necessary upgrades to both online servers and near-line training systems. Therefore, the successful integration of multimodal representations hinges on the ability to boost their performance benefits while concurrently minimizing deployment costs.

To accomplish these two key objectives, three practical challenges should be addressed:

• **Design of the pre-training task.** The effectiveness of multimodal representations in enhancing performance hinges on their ability to provide meaningful semantic information, which is difficult for ID features to capture. It is essential to design pre-training
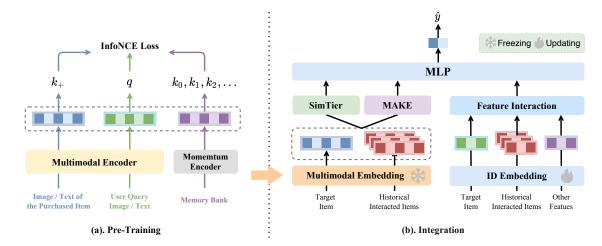
Figure 1: An overview of our two-phase framework: the pre-training of multimodal representations, followed by the integration of pre-trained representations into recommendation models. In the first phase (refer to Figure (a)), we undertake pre-training through semantic-aware contrastive learning. This method equips the multimodal representations with ability to identify semantic similar items. Subsequently, in the second phase (refer to Figure (b)), we introduce our proposed SimTier and make methods to effectively incorporate the pre-trained multimodal representations into the recommendation models.

tasks that enable multimodal representations to encapsulate such semantic information.

- **Integration of multimodal representation.** The inherent discrepancies between ID features and multimodal representations, such as difference in training epochs, calls for approaches that can effectively incorporating multimodal representations into the ID-based model. These approaches should leverage the strengths of each feature type to enhance the model's overall performance.
- **Design of the production system.** The entire production workflow, including the generation of multimodal representations for new items and their up-to-date application in downstream tasks, should be designed with efficiency.

To address these challenges, we adopt a two-phase framework as depicted in Figure 1. Specifically, during the pre-training phase, we propose the **S**emantic-aware **C**ontrastive **L**earning (SCL) method. In this phase, we utilize the user's search query and subsequent purchase action to construct semantically similar sample pairs, capturing the dimensions of semantic similarity that are most relevant to users in e-commerce scenarios. For negative samples, we draw from a large memory bank. The SCL method enables multimodal representations to effectively measure semantic similarities among items.

Upon obtaining high-quality multimodal representations, we propose two approaches to incorporate these representations into the existing ID-based model. Firstly, we develop an approach named SimTier to measure the degree of similarity between the target item and items the user has previously interacted with. The resulting SimTier vector is then concatenated with other embeddings and fed into the subsequent layers. Furthermore, to address the discrepancy in training epochs between multimodal representations and ID embeddings, we introduce the **M**ultimod**A**l **K**nowledge **E**xtractor (MAKE) module. The MAKE module separates the optimization of parameters associated with multimodal representations from

those of the ID-based model, enabling more effective learning for the parameters related to multimodal representations.

We also present the design of our production system that facilitates the deployment of multimodal representations. Specifically, the system generates multimodal representations for newly introduced items in real-time and ensures these representations are immediately available for the training infrastructure and the online prediction server. This design achieves minimal latency—merely a few seconds between the introduction of an item and the model's use of its multimodal representation. Since mid-2023, multimodal representations have been deployed in Taobao display advertising system, leading to significant performance improvements.

2 PRELIMINARIES

Before delving into the specifics, we first introduce the typical ID-based model structure utilized in different stages (including retrieval [15, 38], pre-ranking [27, 34], and ranking stages [1, 3, 7, 10, 36]) of industrial system.

ID Features in Recommendation Models: Common recommendation models are trained on large-scale datasets comprising billions of ID features [16, 32]. These ID features serve to represent users profiles, user historical interacted items, the target item (to be predicted), and the contextual information. For example, we can represent the target item with its respective item ID and category ID and represent user historical interacted items through a sequence of corresponding item IDs and category IDs.

Structure of ID-based Model: The ID-based recommendation model follows an embedding and MLP (Multi-Layer Perceptron) architecture, which typically incorporates *historical behavior modeling* modules [35, 36]. Initially, all ID features are converted into embeddings. The historical behavior modeling modules then measures a user's interest towards the target item by analyzing the relevance between the embedding of target item and embeddings

of the user's historical interacted items. Specifically, the modules produce fixed-length vectors by aggregating the embeddings of the target item and those of the historical interacted items. These vectors are then concatenated with other ID embeddings to form the input for subsequent MLP, which produce the final prediction.

3 PRE-TRAINING OF MULTIMODAL REPRESENTATIONS

As for the multimodal data, its utilization can improve historical behavior modeling. Specifically, multimodal data can be used to measure the semantic similarity between the target item and users' historical interacted items. Take item images as an example, they can be used to measure the visual similarity between the image of the target item and those of historical interacted items. Intuitively, a higher semantic similarity indicates a stronger resemblance between the target item and the users' historical behavior, suggesting a higher likelihood of the user's interest. This insight emphasizes the importance of designing a pre-training task tailored to enable multimodal representations to effectively discern the semantic similarity across item pairs.

3.1 Semantic-Aware Contrastive Learning

To derive representations with the capability to discern semantic similarity, we propose the semantic-aware contrastive learning (SCL) method that attracts the semantically similar sample pairs and repulses the dissimilar sample pairs. To accomplish this, it is essential to define semantically similar and dissimilar pairs for supervision. To understand the importance of this point, consider the example shown in Figure 5, where the three pillows are almost identical yet display slight variances, like differences in patterns or minor appearance discrepancies. If the definition of semantically similar pairs is inadequate, the representation might fail to capture such subtle differences. Indeed, these slight distinctions are frequently missed by representations focused on capturing general concepts.

In the following, we will elaborate our construction of pre-training dataset tailored for e-commerce scenario and the optimization strategies of contrastive learning process.

3.2 Construction of Pre-Training Dataset

In the context of e-commerce, **a user's search query and subsequent purchase action** often signifies a strong semantic similarity between the query and the purchased item. For example, if a user searches for an image of a pillow and subsequently purchases a pillow, this sequence of actions indicates that the two images (the queried image and the image of the purchased item) are semantically similar enough to satisfy the user's purchase intentions. Thus, as shown in Table 1, in training the text encoder, we pair the text of the user's search query with the title of the item they ultimately purchased as the semantically similar pair. Similarly, for the image modality, user's image query (obtained from the image search scenario of Taobao) is paired with the image of the subsequent purchased item. This pairing strategy naturally captures the dimensions of semantic similarity that are most relevant to users in e-commerce scenarios, reflecting the elements that influence their purchasing decisions.

Table 1: The construction of semantically similar pair for pre-training.

modality	semantically similar pair
Image	<user's image query, image of the purchased item>
Text	<user's text query, title of the purchased item>

For each semantically similar pair, we consider all other samples as potential dissimilar samples, which can be achieved by using the samples in the current mini-batch as negatives. To further improve the model performance, we aim to increase the number of negative samples available during training. Specifically, we draw inspiration from MoCo [12] and adopt a technique that updates the model with momentum, facilitating sampling more negatives from a larger memory bank. More sophisticated strategies for constructing negative pairs, such as identifying hard negatives [23], will be elaborated in Section 6.1.2.

3.3 Optimization

We utilize the well-regarded InfoNCE loss [19] as the loss function. Given an encoded query q and its corresponding encoded positive sample k, along with $k_0, k_1, \ldots, k_K$ representing the set of encoded sample representations in the memory bank, where K denotes the memory bank size [12], the InfoNCE loss employs the dot product to measure similarity (with all representations being L2 normalized). As shown in Equation 1, the loss value decreases when query q closely matches its designated positive sample k and diverges from all other samples within the memory bank.

$$L_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot k_+)/\tau}{\sum_{i=0}^{K} \exp(q \cdot k_i)/\tau}. \tag{1}$$

Here, τ represents a learnable temperature parameter. For our experiments, we set the value of K to 196,800.

By this means, the SCL method enables the representations to have the ability to discern fine differences between comparable items, which are crucial for recommendation models.

4 INTEGRATION WITH RECOMMENDATION MODELS

A direct method to integrate multimodal representations into an ID-based recommendation model involves concatenating these multimodal representations with the ID embeddings for both the target item and users' past interacted items [9, 29]. This concatenation is followed by the utilization of user behavior modeling modules, which are subsequently input into the MLP for the final prediction. Although this method is straightforward, we find that the performance improvements are modest. To investigate this matter, we share our observations and insights.

4.1 Observations and Insights

We begin by sharing our observations and insights on incorporating multimodal representations.

Observation 1: simplifying the usage of multimodal representations improves performances. Our research reveals that
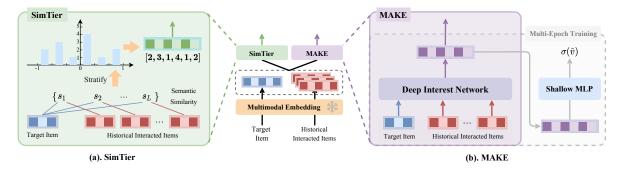
Figure 2: An illustration of the proposed SimTier and Multimodal Knowledge Extractor (MAKE) approaches, with details provided in Section 4.2 and 4.3, respectively.

the direct integration of multimodal representations into the ID-based model does not yield optimal performance, as explained in more detail in Section 6.3. This issue arises because the parameters associated with multimodal representations, e.g., the parameters of the MLP that are connected with multimodal representations, are not adequately learned during the joint training process with the ID embeddings [29]. In contrast, strategies that simplify the usage of multimodal representations, for instance, transforming them into semantic IDs (thereby representing the embedding vectors with IDs) [26, 29], appear to offer improved performance.

Observation 2: ID-based and multimodal-based models have training epoch discrepancy. In industrial scenario, ID-based models are typically trained for only one epoch to avoid overfitting [33]. In contrast, models in CV and NLP field often undergo training over multiple epochs. A natural question arises: how many training epochs are ideal for a multimodal-based recommendation model? To answer this question, we developed a recommendation model that exclusively utilizes multimodal representations as input features, without any ID features, and analyzed how its performance varied with the number of training epochs.

The detailed convergence curve is illustrated in Figure 3. We find that model leveraging multimodal data benefits from training across multiple epochs on the same dataset, with its performance showing notable enhancements as the number of epochs increases. In contrast, ID-based models suffer from the one-epoch overfitting phenomenon, where model performance dramatically degrades at the beginning of the second epoch [33]. The result suggests that the parameters associated with multimodal representations require more epochs to converge properly, which contrasts with the behavior of the ID-based model. **Consequently, when incorporating multimodal representations into an ID-based model, and trained over only one epoch, there is a risk that the multimodal-related parameters may not be sufficiently trained.**

4.2 Method I: SimTier

The observation 1 calls for simplifying the usage of multimodal representations. To this end, we propose a straightforward yet effective method SimTier. As shown in Figure 2 (a), SimTier begins by computing the dot product similarity between the multimodal representation of the target candidate item, denoted as v_c, and the
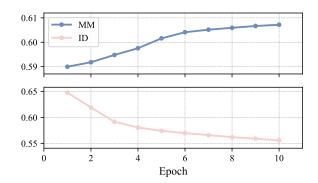

Figure 3: The multimodal-based CTR prediction model (MM) demonstrates a continuous increase in GAUC after several training epochs. In contrast, the ID-based model (ID) shows a sharp decline in GAUC during testing after the second epoch of training.

multimodal representations of the user's historically interacted items $\{v_i\}_{i=1}^{L}$,

$$s_i = v_i \cdot v_c, \forall i \in \{1, \dots, L\}. \tag{2}$$

Following the calculation of the similarity scores, we partition the score range of [-1.0, 1.0] into N predefined tiers. Within each tier, we count the number of similarity scores that fall into that corresponding range. Hence, we obtain an N-dimensional vector, with each dimension representing the number of similarity scores in the corresponding tier. Thus, SimTier effectively converts a set of high-dimensional multimodal representations into a N-dimensional vector that encapsulate the degree of similarity between the target item and the user's historical interactions. The obtained N-dimensional vector is then concatenated with other embeddings and fed to the following MLP. We provide the pseudo code of SimTier in Algorithm 1.

4.3 Method II: Multimodal Knowledge Extractor (MAKE)

To address the difference in training epochs required for ID features versus multimodal representations, we introduce the **M**ultimod**A**l **K**nowledge **E**xtractor (MAKE) module, decoupling the optimization of multimodal related parameters from that of ID features.

Algorithm 1 A Tensorflow-style Pseudocode of the SimTier.

```
# B: batch size, S: sequence length, N: tier number, D: dim
# Input: target [B, D], seq [B, S, D]
# Output: sim_tier [B, N]

# Compute the similarity scores [B, S]
sim_score = reduce_sum(expand_dims(target, 1)*seq, axis=2)
# Assign tier to each score [B, S]
indices = reshape(ceil((sim_score + 1) / 2 * N), [-1, S])
# Accumulate counts for each tier [B, N]
weight = equal(
        reshape(range(0, N, 1), [1, N, 1]),
        expand_dims(indices, axis=1)
    )
sim_tier = reduce_sum(weight, axis=2, keep_dims=True)
```

The MAKE module consist of two steps: 1) multi-epoch training to extract useful multimodal knowledge and 2) knowledge utilization by the downstream task.

Multi-epoch training of multimodal related parameters. The goal of the MAKE module is pre-training the parameters related to multimodal representations over multiple epochs to ensure their convergence. In practice, we utilize the CTR prediction task as the recommendation pre-training task. As shown in Figure 2 (b), we first develop a DIN-based user behavior modeling module [36]. This module processes the pre-trained multimodal representations of the target item and historical interacted items, resulting in the output $\mathbf{v}_{\text{MAKE}}$:

$$\mathbf{v}_{\text{MAKE}} = \text{DIN}(\{\mathbf{v_i}\}_{\mathbf{i=0}}^{\text{L}}, \mathbf{v_c}). \tag{3}$$

After that, $\mathbf{v}_{\text{MAKE}}$ is fed into a four-layer Multi-layer Perceptron (MLP_{MAKE}) and produced the logit $\hat{v}$

$$\hat{v} = \text{MLP}_{\text{MAKE}}(\mathbf{v}_{\text{MAKE}}). \tag{4}$$

Then we optimize the cross-entropy loss between the predicted click probability and the binary click label y, as shown in Equation 5.

$$\mathcal{L}_{\text{MAKE}} = \sum -y \log \sigma(\hat{v}) - (1 - y) \log(1 - \sigma(\hat{v})) \tag{5}$$

The recommendation pre-training task allows the MAKE module to refine its parameters via training over multiple epochs and extract knowledge $\mathbf{v}_{\text{MAKE}}$ from multimodal representations, thereby enhancing its effectiveness for recommendation tasks.

Knowledge utilization. After acquiring the vector $\mathbf{v}_{\text{MAKE}}$, the subsequent step is to integrate it into the downstream recommendation task. Practically, we concatenate $\mathbf{v}_{\text{MAKE}}$ and intermediate outputs from MLP_{MAKE} with other embeddings and input this combined data into the subsequent layers. The implementation of the multi-epoch training of MAKE module reconciles the training epochs difference required for ID embeddings and multimodal representations, resulting in better performances.

5 INDUSTRIAL DESIGN FOR ONLINE DEPLOYMENT

In industrial settings, new items (including advertisements) are constantly being created. To maintain prediction accuracy for these new items, it's crucial for the recommendation model to acquire the multimodal representation of new items in real-time. This calls for the ability of continuous generating multimodal representations for new items and a real-time utilization by near-line trainer and
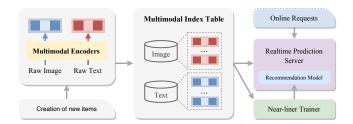

Figure 4: An overview of the online system.

online server. The illustrative overview of our online system is provided in Figure 4. To achieve the real-time generation of multimodal representations, upon the introduction of new items, the system automatically initiates a request to the pre-trained multimodal encoders to compute the multimodal representations for these new items. Once inferred, these representations are sent to the multimodal index table. Following this step, the downstream training systems and inference servers are able to retrieve the multimodal representations from the index table, facilitating near-line training and real-time online prediction capabilities. This process ensures minimal latency—reduced to just a few seconds—between an item's introduction and the utilization of its multimodal representation by the model.

6 EXPERIMENT

In this section, we delve into a case study that examines the integration of image representations into the CTR prediction model, aiming to provide a comprehensive analysis. It is notable that our approach is general, capable of accommodating several modal types (such as text and video) and applicable across different stages within the recommendation systems.

6.1 Experiment Setup

6.1.1 Datasets. We start by detailing the datasets used for the pre-training phase and the downstream integration phase.

- **Pre-training Dataset.** In the pre-training dataset, each sample consists of a Query (user's image query) and a Positive (the image of the purchased item). To further enhance the performance, we also add a hard Negative (the image of the clicked item triggered by the positive item). A case of the pre-training dataset is depicted in Figure 5.
- **CTR Prediction Dataset.** The CTR prediction dataset is obtained from the Taobao display advertising system, using impression logs of one week.

6.1.2 Compared Pre-training Methods. We employ a range of widely-used pre-training method for comparison.

- **CLIP-O.** CLIP-O refers to the CLIP visual encoder (CLIP-ViT-B/16) that has been pre-trained using a universal dataset [21].
- **CLIP-E.** CLIP-E is the fine-tuned version based on the CLIP-O model in the e-commerce scenario using aligned item descriptions and item images.
- **SCL.** The proposed semantic-aware pre-training method. The SCL approach employs Momentum Contrast (MoCo) to expand the set of negative samples. Furthermore, SCL applies triplet

Figure 5: A case of the pre-training dataset.

loss [23] to each <Query, Positive, Negative> triplet to effectively discriminate hard negatives. Our experimental analysis investigate the impact of excluding triplet loss and MoCo to assess their contributions to the overall performance.

6.1.3 Compared Recommendation Methods. We employ a range of widely-used integration approach for comparison.

- ID-based Model (production baseline). The baseline is a ID-based model that underpins our online system.
- Vector. The Vector method utilize the pre-trained multimodal representations as the side-information of each item, and concatenated them with other ID embeddings within the model.
- SimScore. Similarity Score (SimScore) can be seen as a simplified version of the Vector method. The semantic similarity score for each historical interacted item with respect to the target item is used as side information.
- SIMTIER and MAKE. The proposed SIMTIER and MAKE approaches.

6.1.4 Evaluation Metrics. We evaluate both the pre-training performance and CTR prediction performance of the proposed methods.

- **Evaluating Pre-training Methods.** Throughout our extensive experimentation, we discovered that the Top-N accuracy (Acc@N) is well correlated with the performance of downstream recommendation models. In detail, the Acc@N metric quantifies the ability of the representation to identify semantic similar items:

$$\text{Acc@N} = \frac{1}{D} \sum_{i=1}^{D} \mathbb{I}(p_i \in \text{Top}_N(q_i, S)), \quad (6)$$

where D denotes the size of test set. The terms q_i and p_i correspond to the query and the positive of the i-th sample, respectively. $S = \{pos_i\}_{i=1}^{D}$ represents the set comprising all positives, and Top_N is a function that retrieves the top-N results for each query by leveraging a multimodal representation from the set S. The symbol $\mathbb{I}(\cdot)$ signifies an indicator function that yields a value of 1 when the i-th p_i is among the retrieval results for q_i, and 0 in all other cases.

- **Evaluating Recommendation Methods.** We assess the effectiveness of the CTR prediction model using the AUC and Group AUC (GAUC) metrics, where a higher AUC/GAUC value signifies superior ranking ability [25, 36, 37].

6.2 Performance on Pre-Training Dataset

6.2.1 Rationale for Utilizing Accuracy to Evaluate the Effectiveness of Pre-trained Representations. The most precise way to gauge the effectiveness of pre-trained representations is by measuring the improvement in recommendation accuracy with the integration of

Table 2: Pre-training performance of different methods

Method	Acc@1	Acc@5
CLIP-O	0.2559	0.3575
CLIP-E	0.2952	0.3917
SCL	**0.7474**	**0.8850**
w/o Triplet Loss	0.6957	0.8604
w/o Triplet Loss & MoCo	0.5760	0.7590

Table 3: Overall performance on CTR prediction dataset.

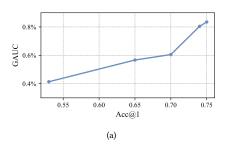| Method	GAUC	AUC
ID-based Model	-	-
Vector	+0.29%	+0.18%
SimScore	+0.77%	+0.40%
SIMTIER	+0.96%	+0.59%
MAKE	+0.93%	+0.51%
SIMTIER+MAKE	**+1.25%**	**+0.75%**

multimodal representations. However, this evaluation process can be lengthy for iteration of pre-training methods, and an intermediary metric for a quicker assessment of pre-trained multimodal representations is desirable.
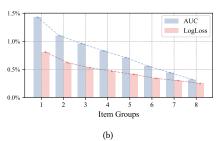
In our research, we observed **a strong correlation between the enhancement in pre-training accuracy and the boost in recommendation performance**. We illustrate this relationship in Figure 6(a), where we can see that the improvement of Acc@1 is consistent with the improvement of GAUC. Hence, we predominantly rely on pre-training accuracy to determine the quality of multimodal representations. The exploration of other potential intermediary metrics for evaluating pre-trained multimodal representations remains a interesting topic for future work.

6.2.2 Importance of Semantic-Aware Contrastive Learning. To investigate how different pre-training tasks affect the quality of multimodal representations, we conducted a series of experiments. The results, presented in Table 2, offer two important observations. First, the proposed SCL pre-training method surpasses other semantic-similarity-agnostic methods, emphasizing **the necessity of the semantic-aware learning.**. Second, incorporating techniques like Momentum Contrast (MoCo) [12] and Triplet Loss [23] further enhances the quality of the multimodal representations, demonstrating the choice of negative sample greatly impacts the performance.

6.3 Performance on CTR prediction Dataset

6.3.1 Performance on Different Integration Strategies. In the CTR prediction dataset, we evaluate the proposed SIMTIER and MAKE against other methods. The overall results are shown in Table 3, from which two observations can be noted. Firstly, SIMTIER and MAKE outperform other methods significantly. Secondly, the combination of SIMTIER and MAKE can further improve the performance, with a 1.25% increase in GAUC, 0.75% increase in AUC compared
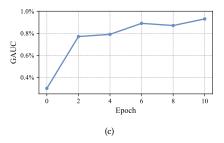
Figure 6: (a) The correlation between pre-training metric and CTR prediction performance. (b) The relative improvement on item groups with different frequency. (c) The relative performance of the recommendation model with different pre-training epochs on MAKE**. Note that 0 epoch implies that** MAKE **does not undergo pre-training, but is instead jointly optimized with the downstream task.**

with the ID-based model. The above results demonstrate the effectiveness of the proposed methods on integrating multimodal representations into the ID-based model.

6.3.2 Performance on Different Training Epochs of MAKE. To explore the impact of multi-epoch pre-training of MAKE on the final recommendation performance, we conduct experiments where MAKE is pre-trained for varying numbers of epochs before integration into the recommendation model. The results are shown in Figure 6(c). Note that 0 epochs implies that MAKE does not undergo pre-training and is instead jointly optimized with the downstream task. The results indicate that as the number of pre-training epochs for MAKE increases, the performance of the final recommendation model also improves. This demonstrates that multi-epoch training of MAKE effectively enhances model performance.

6.3.3 Performance on Infrequent Items. To examine the generalization ability of multimodal representation on long-tail items, we assess the relative improvement across item groups categorized by their frequency of occurrence. We divide all items into eight groups, with Group 1 containing items of the lowest frequency and Group 8 encompassing those with the highest frequency in the training dataset. Afterward, we compute the relative improvement of AUC as $|AUC_{MM}\text{-}AUC_{ID}|/AUC_{ID}$ for each group, where *MM* denotes the combination of SimTier and MAKE method and *ID* represents the ID-based production baseline model. We also compute the relative improvement of LogLoss as $|LogLoss_{MM}\text{-}LogLoss_{ID}|/LogLoss_{ID}$ to measure the calibration ability [11, 24]. The result presented in the Figure 6(b) indicates that the multimodal representation exhibits a significant improvement in all groups, demonstrating the effectiveness of our model over different types of items. Meanwhile, we see **a more significant improvement for low-frequency items**. The result demonstrate that multimodal representations can address the shortcomings of ID-based model on long-tail items and enhance the prediction accuracy.

6.4 Online Performance

Since mid-2023, multimodal representations have been integrated into pre-ranking, ranking, and re-ranking models within the Taobao display advertising system, resulting in substantial performance improvements. For instance, incorporating image representations

in the CTR prediction model yielded a overall 3.5% increase in CTR, a 1.5% boost in RPM, and a 2.9% rise in ROI. Notably, the impact was even more pronounced for new ads (created within the last 24 hours), with improvements of 6.9% in CTR, 3.7% in RPM, and 7.7% in ROI. The significant gains on new ads also validate the effectiveness of multimodal data in mitigating the cold-start issue.

7 RELATED WORK

Currently, ID features constitute the core of industrial recommendation models [2, 5, 7, 14, 31, 33, 35, 36]. Despite their widespread adoption, ID features have notable limitations, including the challenge of capturing semantic information and the persistent issue of the cold-start problem [9, 18, 22, 28]. In contrast, multimodal data offer rich semantic information, prompting numerous studies to explore their incorporation into recommendation models. Some research has investigated the potential of learning multimodal representations in an end-to-end manner alongside recommendation model training [4, 8, 30]. However, the substantial computational resources required for such processes often preclude their adoption in industrial systems. Therefore, our focus is on the two-phase paradigm [6, 9, 13, 17, 20, 26, 29], and we aim to share our methods and the valuable insights.

8 CONCLUSION AND DISCUSSION

Multimodal-based recommendation has attracted attention over decades. However, integrating multimodal representations into industrial systems presents many hard challenges, particularly in the realms of representation quality, integration methods, and system implementation—challenges that are amplified within the context of large-scale industrial systems.

In this study, we delve into these challenges and share the approaches we employed for pre-training and incorporation of multimodal representations. Additionally, we provide insights gleaned from our experiences during the online deployment stage. We believe the strategies and insights we have amassed through our journey will serve as a valuable resource for those aiming to expedite the adoption of multimodal-based recommendations in industrial systems.

REFERENCES

[1] Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Yong-Nan Zhu, Zhangming Chan, Na Mou, Xinchen Luo, Shiming Xiang, Guorui Zhou, Xiaoqiang Zhu, and Hongbo Deng. 2022. CAN: Feature Co-Action Network for Click-Through Rate Prediction. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. 57–65.

[2] Zhangming Chan, Yuchi Zhang, Xiuying Chen, Shen Gao, Zhiqiang Zhang, Dongyan Zhao, and Rui Yan. 2020. Selection and generation: Learning towards multi-product advertisement post generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3818–3829.

[3] Zhangming Chan, Yu Zhang, Shuguang Han, Yong Bai, Xiang-Rong Sheng, Siyuan Lou, Jiacen Hu, Baolin Liu, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. Capturing Conversion Rate Fluctuation during Sales Promotions: A Novel Historical Data Reuse Approach. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3774–3784.

[4] Junxuan Chen, Baigui Sun, Hao Li, Hongtao Lu, and Xian-Sheng Hua. 2016. Deep CTR Prediction in Display Advertising. In *Proceedings of the 2016 ACM Conference on Multimedia Conference*. 811–820.

[5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. ACM, 7–10.

[6] Haibin Cheng, Roelof van Zwol, Javad Azimi, Eren Manavoglu, Ruofei Zhang, Yang Zhou, and Vidhya Navalpakkam. 2012. Multimedia features for click prediction of new ads in display advertising. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 777–785.

[7] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 191–198.

[8] Shereen Elsayed, Lukas Brinkmeyer, and Lars Schmidt-Thieme. 2022. End-to-End Image-Based Fashion Recommendation. *CoRR* abs/2205.02923 (2022).

[9] Tiezheng Ge, Liqin Zhao, Guorui Zhou, Keyu Chen, Shuying Liu, Huiming Yi, Zelin Hu, Bochao Liu, Peng Sun, Haoyu Liu, Pengtao Yi, Sui Huang, Zhiqiang Zhang, Xiaoqiang Zhu, Yu Zhang, and Kun Gai. 2018. Image Matters: Visually Modeling User Behaviors Using Advanced Model Server. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2087–2095.

[10] Siyu Gu, Xiang-Rong Sheng, Ying Fan, Guorui Zhou, and Xiaoqiang Zhu. 2021. Real Negatives Matter: Continuous Training with Real Negatives for Delayed Feedback Modeling. In *Proceedings of The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2890–2898.

[11] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. 1321–1330.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9726–9735.

[13] Ruining He and Julian J. McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web*. 507–517.

[14] Jiacen Hu, Zhangming Chan, Yu Zhang, Shuguang Han, Siyuan Lou, Baolin Liu, Han Zhu, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. PS-SA: An Efficient Self-Attention via Progressive Sampling for User Behavior Sequence Modeling. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4639–4645.

[15] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based Retrieval in Facebook Search. In *Proceedings of The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2553–2561.

[16] Biye Jiang, Chao Deng, Huimin Yi, Zelin Hu, Guorui Zhou, Yang Zheng, Sui Huang, Xinyang Guo, Dongyue Wang, Yue Song, et al. 2019. XDL: An Industrial Deep Learning Framework for High-Dimensional Sparse Data. In *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*. 1–9.

[17] Corey Lynch, Kamelia Aryafar, and Josh Attenberg. 2016. Images Don't Lie: Transferring Deep Visual Semantic Features to Large-Scale Multimodal Learning to Rank. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 541–548.

[18] 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 Twenty-Fourth International Joint Conference on Artificial Intelligence*. 3728–3734.

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[20] Aditya Pal, Chantat Eksombatchai, Yitong Zhou, Bo Zhao, Charles Rosenberg, and Jure Leskovec. 2020. PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest. In *Proceedings of The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2311–2320.

[21] 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. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. 8748–8763.

[22] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 253–260.

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 815–823.

[24] Xiang-Rong Sheng, Jingyue Gao, Yueyao Cheng, Siran Yang, Shuguang Han, Hongbo Deng, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. Joint Optimization of Ranking and Calibration with Contextualized Hybrid Model. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4813–4822.

[25] Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, and Xiaoqiang Zhu. 2021. One Model to Serve All: Star Topology Adaptive Recommender for Multi-Domain CTR Prediction. In *Proceedings of The 30th ACM International Conference on Information and Knowledge Management*. 4104–4113.

[26] Anima Singh, Trung Vu, Raghunandan H. Keshavan, Nikhil Mehta, Xinyang Yi, Lichan Hong, Lukasz Heldt, Li Wei, Ed H. Chi, and Maheswaran Sathiamoorthy. 2023. Better Generalization with Semantic IDs: A case study in Ranking for Recommendations. *CoRR* abs/2306.08121 (2023).

[27] 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).

[28] Kailun Wu, Weijie Bian, Zhangming Chan, Lejian Ren, Shiming Xiang, Shu-Guang Han, Hongbo Deng, and Bo Zheng. 2022. Adversarial gradient driven exploration for deep click-through rate prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2050–2058.

[29] Jia-Qi Yang, Chenglei Dai, Dan Ou, Ju Huang, De-Chuan Zhan, Qingwen Liu, Xiaoyi Zeng, and Yang Yang. 2023. COURIER: Contrastive User Intention Reconstruction for Large-Scale Pre-Train of Image Features. *CoRR* abs/2306.05001 (2023).

[30] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to Go Next for Recommender Systems? ID-vs. Modality-based Recommender Models Revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2639–2649.

[31] Yujing Zhang, Zhangming Chan, Shuhao Xu, Weijie Bian, Shuguang Han, Hongbo Deng, and Bo Zheng. 2022. KEEP: An Industrial Pre-Training Framework for Online Recommendation via Knowledge Extraction and Plugging. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3684–3693.

[32] Yuanxing Zhang, Langshi Chen, Siran Yang, Man Yuan, Huimin Yi, et al. 2022. PICASSO: Unleashing the Potential of GPU-centric Training for Wide-and-deep Recommender Systems. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. 3453–3466.

[33] Zhao-Yu Zhang, Xiang-Rong Sheng, Yujing Zhang, Biye Jiang, Shuguang Han, Hongbo Deng, and Bo Zheng. 2022. Towards Understanding the Overfitting Phenomenon of Deep Click-Through Rate Models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2671–2680.

[34] Zhishan Zhao, Jingyue Gao, Yu Zhang, Shuguang Han, Siyuan Lou, Xiang-Rong Sheng, Zhe Wang, Han Zhu, Yuning Jiang, Jian Xu, and Bo Zheng. 2023. COPR: Consistency-Oriented Pre-Ranking for Online Advertising. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4974–4980.

[35] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep Interest Evolution Network for Click-Through Rate Prediction. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. Honolulu, Hawaii, USA, 5941–5948.

[36] 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*. ACM, 1059–1068.

[37] Han Zhu, Junqi Jin, Chang Tan, Fei Pan, Yifan Zeng, Han Li, and Kun Gai. 2017. Optimized Cost per Click in Taobao Display Advertising. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2191–2200.

[38] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. 2018. Learning Tree-based Deep Model for Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London, UK, 1079–1088.