SEMI: A Sequential Multi-Modal Information Transfer Network for E-Commerce Micro-Video Recommendations

Chenyi Lei
leichy@mail.ustc.edu.cn
University of Science and Technology of China & Alibaba Group
Hefei, Anhui, China

Yong Liu*
stephenliu@ntu.edu.sg
Nanyang Technological University
Singapore

Lingzi Zhang
lingzi001@e.ntu.edu.sg
Nanyang Technological University & Alibaba Group
Singapore

Guoxin Wang
xiaogong.wgx@taobao.com
Zhejiang University & Alibaba Group
Hangzhou, Zhejiang, China

Haihong Tang
piaoxue@taobao.com
Alibaba Group
Hangzhou, Zhejiang, China

Houqiang Li*
lihq@ustc.edu.cn
University of Science and Technology of China
Hefei, Anhui, China

Chunyan Miao
ascymiao@ntu.edu.sg
Nanyang Technological University
Singapore

ABSTRACT

The micro-video recommendation system becomes an essential part of the e-commerce platform, which helps disseminate micro-videos to potentially interested users. Existing micro-video recommendation methods only focus on users' browsing behaviors on micro-videos, but ignore their purchasing intentions in the e-commerce environment. Thus, they usually achieve unsatisfied e-commerce micro-video recommendation performances. To address this problem, we design a **SE**quential **M**ulti-modal **I**nformation transfer network (SEMI), which utilizes product-domain user behaviors to assist micro-video recommendations. SEMI effectively selects relevant items (*i.e.,* micro-videos and products) with multi-modal features in the micro-video domain and product domain to characterize users' preferences. Moreover, we also propose a **C**ross-domain **C**ontrastive **L**earning (CCL) algorithm to pre-train sequence encoders for modeling users' sequential behaviors in these two domains. The objective of CCL is to maximize a lower bound of the mutual information between different domains. We have performed extensive experiments on a large-scale dataset collected from Taobao, a world-leading e-commerce platform. Experimental results show that the proposed method achieves significant improvements over state-of-the-art recommendation methods. Moreover, the proposed method has also been deployed on Taobao, and the online A/B testing results further demonstrate its practical value.

*Corresponding Author.

KDD '21, August 14–18, 2021, Virtual Event, Singapore.

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467189

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

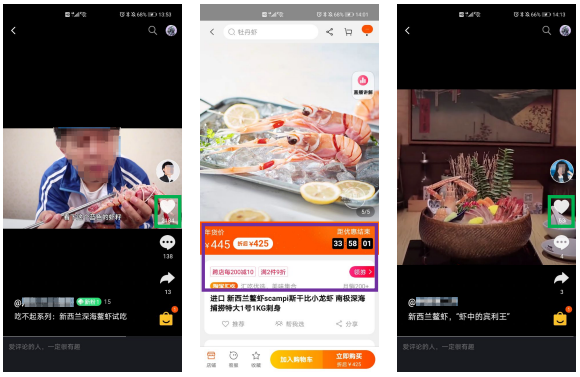
KEYWORDS

Micro-video recommendation; cross-domain recommendation; sequential recommendation

ACM Reference Format:
Chenyi Lei, Yong Liu, Lingzi Zhang, Guoxin Wang, Haihong Tang, Houqiang Li, and Chunyan Miao. 2021. SEMI: A Sequential Multi-Modal Information Transfer Network for E-Commerce Micro-Video Recommendations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3447548.3467189

1 INTRODUCTION

Micro-videos are an emerging type of content on e-commerce platforms. They focus on introducing merchandise in various video display forms, *e.g.,* opinion leaders' analyses and syntheses, product experience sharing from grass-root users, and sitcom advertisements crafted by product providers or merchants. For e-commerce platforms, micro-videos have the following crucial advantages. Firstly, providing users with high-quality micro-videos enables different pan-entertainment experiences, which significantly increases the user stickiness to the platform. Secondly, well-created micro-videos would attract users' interests in the related merchandise and inspire their shopping behaviors. Thirdly, micro-videos provide a high-quality channel for merchants to perform brand promotion and online marketing. On Taobao, a world-leading e-commerce platform, there are more than 100 million micro-videos, and new micro-videos are generated at the speed of hundreds of thousands per day. More than 30 million users browse micro-videos every day, which leads to an estimated gross merchandise volume of million dollars. Accordingly, there is a great need for e-commerce providers



(1) The Micro-Video (2) The Product (3) The Micro-Video

Figure 1: Screenshots of the Taobao mobile app. (2) shows the product page about crawfish, which consists of a series of promotion information (purple box). (1) and (3) are two micro-videos about crawfish, where (1) is more popular than (2) according to the number of "likes" (green box).

to build effective micro-video recommendation systems to enhance the user experience, engagement, and retention.

Compared to videos on traditional online video sharing platforms, *e.g.,* YouTube and Youku, micro-videos usually have a much shorter length, *e.g.,* tens of seconds. A user can interact with more micro-videos at a lower decision cost. Therefore, users' interaction behaviors with micro-videos usually follow more dynamic patterns than their behaviors with traditional online videos. To a certain extent, this makes micro-video recommendation [8, 16, 21, 22, 34] more challenging than traditional video recommendation task [10, 11, 32]. On conventional micro-video sharing platforms, *e.g.,* Kuaishou, TikTok, and Vine, there are no constraints on the topics of micro-videos. However, the micro-videos on e-commerce platforms are mainly about products. The users' behavioral intentions on these platforms are usually different. For example, on TikTok, the users interact with micro-videos mainly for having fun. When interacting with micro-videos on Taobao, besides entertainment, the users are more prepared to be "inspired" or purchase a specific kind of products. Therefore, existing video and micro-video recommendation methods usually achieve unsatisfied performances for micro-video recommendation on e-commerce platforms without considering the users' purchasing intentions.

On e-commerce platforms, there is a large volume of user behavior data on products, which can help identify users' preferences and purchasing intentions. In this work, we propose to leverage the product-related user behavior data for micro-video recommendation. It is a very challenging task, because there exists a large gap between users' behaviors on products and micro-videos. Although micro-videos are about products, their forms are quite different, which is the crucial reason influencing users' behavior patterns in these two domains [33]. Products are well structured with clearly defined attributes, *e.g.,* price and keyword descriptions. However, micro-videos are the multimedia content created based on product information. Besides the introduced products, whether a user enjoys a micro-video is more dependent on her tastes on the multi-modal information of the micro-video. For better understanding, we illuminate an example in Figure 1, where all three sub-figures describe the New Zealand crawfish. As shown in Figure 1 (2), the product page is structured with promotion information, clean cover pictures, and keyword-like titles, making users more inclined to purchase the product directly and quickly. On the contrary, micro-videos are more diverse with multimedia content. The video in Figure 1 (1) tends to attract users through a humorous tasting topic, while the video in Figure 1 (3) shows more ingredients in high-end restaurants. In addition, different multi-modal information display forms of micro-videos may also cause different popularity.

This study focuses on developing domain transfer and adaptation techniques to effectively exploit users' sequential behaviors in product-domain for sequential micro-video recommendation. We propose a novel cross-domain recommendation method, named **SE**quential **M**ulti-modal **I**nformation transfer network (SEMI) for our task. Specifically, SEMI learns users' cross-domain interests to overcome the differences of content structures and users' sequential behavior patterns in different domains. Hierarchical transformer structure [31] is employed to analyze the intra-relationship between behaviors in each single domain and the inter-relationship between cross-domain behaviors. To better bridge the behavior pattern gap between micro-video domain and product domain, we further propose a **C**ross-domain **C**ontrastive **L**earning (CCL) algorithm to pre-train encoders to describe sequential user behaviors. CCL imposes dual constraints on the cross-domain sequential behaviors by minimizing the gap between micro-videos and products' representations in the same user behavior sessions and maximizing the distances between randomly sampled pairs.

The main contributions made in this work are as follows:

- To the best of our knowledge, this is the first study investigating e-commerce micro-video recommendation, whose importances and business values have been discussed before.
- We propose to use product-domain user behaviors to enhance the micro-video recommendation performances. To that end, we propose the **SE**quential **M**ulti-modal **I**nformation transfer network (SEMI), which effectively selects relevant items in both domains to characterize user preferences.
- To bridge the gap between micro-video domain and product domain, we propose a **C**ross-domain **C**ontrastive **L**earning (CCL) algorithm to pre-train encoders for modeling users' sequential behaviors from the multimodality view. CCL-based pre-training strategy significantly improves the performances of SEMI.
- We conduct extensive experiments on a large-scale dataset collected from Taobao. The experimental results verify that the proposed method outperforms state-of-the-art recommendation methods. We have also deployed the proposed method on Taobao, and online A/B testing results demonstrate its practical values.

2 RELATED WORK

2.1 Video Recommendation

Existing video recommendation methods can be divided into three main categories, *i.e.,* collaborative filtering based methods [2, 35], content-based methods [6, 10, 32], and hybrid approaches [3, 4]. These approaches mainly deal with traditional online videos instead of micro-videos. With the growing popularity of micro-videos, micro-video recommendation has attracted increasing research attentions [7, 8, 16, 19, 21, 22, 26, 34]. Compared with traditional

video recommendation task, micro-video recommendation needs more targeted research to address its unique challenges, *i.e.,* diverse and dynamic user interests. Existing micro-video recommendation methods mainly focus on exploiting the multi-modal features of micro-videos and the users' sequential behaviors. For example, [8] proposes the temporal hierarchical attention at category- and item-level network for the micro-video recommendation. [22] designs a user-video co-attention network to learn multi-modal information from both users and micro-videos for recommendation. [16] proposes an end-to-end multi-scale time-aware user interest modeling network to explore multi-scale time effects on the user interests. Although these studies usually achieve promising performance improvements, they cannot be directly applied to the micro-video recommendation on e-commerce platforms. Because they only focus on the user-video interactions without considering the users' purchasing intentions.

2.2 E-Commerce Recommendation

Personalized recommendation systems have been widely deployed on e-commerce platforms [5, 12, 14, 18, 20, 23, 30, 38, 39]. For example, [39] proposes the Deep Interest Network that learns the user representations from historical behaviors adaptively. [30, 38] further investigate the users' dynamic and long-term interests and propose the Deep Interest Evolution Network and Multi-channel user Interest Memory Network, respectively. [20] studies the multi-task problem in e-commerce platforms and proposes a Pareto-efficient algorithm. Previous studies about e-commerce recommendation tasks mainly focus on recommending products and investigating users' sequential behaviors on products. To the best of our knowledge, none of them explores the micro-video recommendation on e-commerce platforms. As most micro-videos on e-commerce platforms are about products, and the e-commerce environment determines users' overall shopping mentality, it is natural to exploit users' behaviors on products to enhance the micro-video recommendation. Nevertheless, micro-videos have unique characteristics different from products, *e.g.,* diverse display forms and different users' behavior patterns. Thus, the micro-video recommendation on e-commerce platforms is still very challenging.

2.3 Cross-Domain Recommendation

Previous cross-domain recommendation approaches can be broadly categorized into two groups, *i.e.,* traditional methods and deep learning-based methods. Traditional methods transfer knowledge from source domain using shallow models. For example, [24] utilizes several matrix factorization models separately on different domains to model user preferences. [15] proposes to use tensor-based factorization to share latent features between different domains. Compared with traditional methods, deep learning-based methods empoly neural networks to learn high-level and complex information within multiple domains [1, 13, 27–29, 36, 37]. For example, [1] introduces a collaborative cross-network to enable dual knowledge transfer across different domains. [13] proposes a multi-view deep learning framework to recommend items by combining rich user features from different domains. [29] transfers users' interests from the news domain to assist advertisement recommendation by the mixed interest network. Moreover, [27] investigates how to share the information from different domains of the same user account and proposes a parallel information-sharing sequential recommendation model. In summary, none of these cross-domain recommendation methods are designed for e-commerce micro-video recommendation, and they lack the exploration of how to maximize mutual information of sequential behaviors with multi-modal view.

3 PROPOSED RECOMMENDATION MODEL

Figure 2 shows the proposed cross-domain e-commerce micro-video recommendation framework. In this section, we first introduce the problem formulation. Then, we introduce some basic components of the proposed framework. Next, we introduce the proposed CCL algorithm and the details of SEMI.

3.1 Problem Formulation

The most important step of personalized e-commerce micro-video recommendation is to estimate the probability that a user would like to watch a specific micro-video. Particularly, in this work, the "watch" behavior stands for watching the micro-video completely. As the user's interests are dynamic, we use her recent behaviors to help predict. Here, a user's recent behaviors include her recent clicking behaviors on products and recent watching behaviors on micro-videos. For simplicity, we omit other personal data of users, such as demographics.

We denote the target micro-video by v_t, the user's product clicking sequence and micro-video watching sequence by $P = \{p_1, p_2, \cdots, p_l\}$ and $V = \{v_1, v_2, \cdots, v_n\}$, respectively, where l and n denote the length of each sequence. The prediction problem can then be formalized as follows,

$$Prob(v_t|P, V) \sim f(P, V, v_t), \tag{1}$$

where the user is represented by her recently interacted micro-videos V and products P, and $Prob(v_t|P, V)$ is the probability that the user would like to watch the target micro-video v_t. Moreover, $f(P, V, v_t)$ is the model used to estimate $Prob(v_t|P, V)$.

3.2 Basic Components

The proposed recommendation model includes the following three basic components, *i.e.,* feature extractor, multi-head attention block, and multi-modal sequence encoder.

3.2.1 Feature Extractor. In this work, we extract multi-modal features for micro-videos and products, which include micro-video titles, micro-video tags, micro-video frames, product titles, product covers, and product categories. Given the multi-modal information of micro-videos and products, the next step is to obtain their semantic representations. Inspired by the powerful information fusion ability of large-scale multi-modal pre-training models [9, 25], we adopt UniVL [25] to fuse the multi-modal information of micro-videos. Specifically, we utilize micro-video titles and frames as inputs and pre-train the UniVL model using about 100 million e-commerce micro-videos. The pre-training tasks of UniVL include masked language modeling, masked frame modeling, video-text alignment, and tag classifications. Finally, for each micro-video v, we have its multi-modal representation $v^m \in \mathcal{R}^{128}$, which describes the semantic information of micro-video frames, titles, and tags. Similarly, we have the semantic representation $p^m \in \mathcal{R}^{128}$

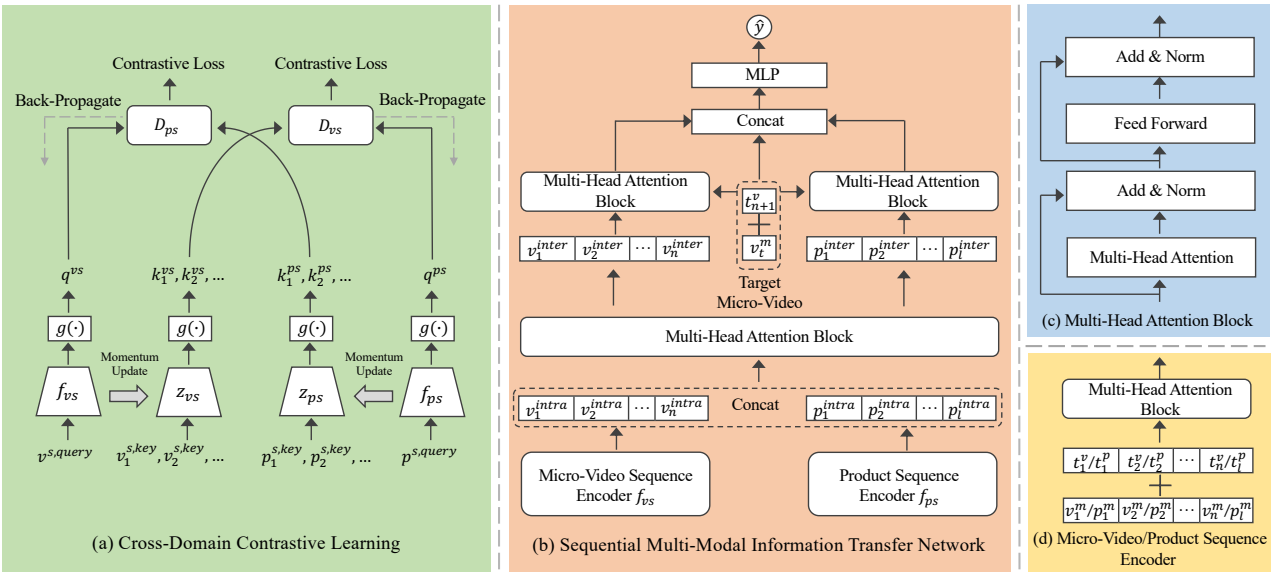


Figure 2: The framework of our proposed method (best viewed in color). Parts (c) and (d) are two basic structures used in our method (c.f. Section 3.2). Part (a) illustrates the proposed CCL algorithm used to pre-train sequence encoders for different domains (c.f. Section 3.3). Part (b) describes the proposed SEMI transfer network for e-commerce micro-video recommendation (c.f. Section 3.4), where the sequence encoders of two domains are initialized by parameters pre-trained by CCL.

for each product p, which is extracted by the pre-trained UNITER model [9]. It describes the semantic information of product cover pictures, titles, and categories. In this work, the UNITER model is pre-trained by about 100 million products. More details about our feature extractors can be found in Appendix A.5.

3.2.2 Multi-Head Attention Block. Motivated by the success of multi-head attention mechanism [31] in sequential recommendation [5], we apply it to capture users' sequential behavior patterns on micro-videos and products. The structure of the attention mechanism module is shown in Figure 2 (c). Following [31], we linearly project an input vector n_h times (*i.e.,* using n_h-heads) to latent vectors with d_h dimensions, by different linear projections. Next, we perform scaled dot-product attention for each head and concatenate all scaled dot-product attention output vectors as follows,

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head_1, head_2, \cdots, head_{n_h}), \\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V), \end{aligned} \tag{2}$$

where $Q \in \mathcal{R}^d$, $K \in \mathcal{R}^d$, and $V \in \mathcal{R}^d$ represent Query, Key, and Value, respectively. $W_i^Q \in \mathcal{R}^{d \times d_h}$, $W_i^K \in \mathcal{R}^{d \times d_h}$, and $W_i^V \in \mathcal{R}^{d \times d_h}$ are weight matrices. The scaled dot-product attention function is defined as follows,

$$Attention(Q,K,V) = softmax(\frac{QK^\top}{\sqrt{d_k}}), \tag{3}$$

where d_k is the dimension of K and equals d_h in our work. Then, the multi-head attention network employs a feed-forward network for further strengthening the model performances. The feed-forward network includes two full-connection layers, and it is applied to each input vector separately and identically.

3.2.3 Multi-Modal Sequence Encoder. Figure 2 (d) shows the structure of the micro-video and product sequence encoders. As the multi-head attention mechanism is not aware of the sequence order information, we use timestamp embedding to preserve the sequential order information for capturing dynamic user preferences over time. Firstly, we split the user's historical behaviors into multiple time blocks. The length of each time block is empirically set to one hour. Then, we learn a timestamp embedding for each time block, which is denoted as $t_i^v \in \mathcal{R}^{128}$ and $t_i^p \in \mathcal{R}^{128}$ for micro-videos and products, respectively. Here, t_i^v and t_i^p share the embedding for the same time block. The input feature of each micro-video $v_i^{in} \in \mathcal{R}^{128}$ ($i \in [1, n]$) is derived by summing the pre-extracted multi-modal representation $v_i^m \in \mathcal{R}^{128}$ and the timestamp embedding $t_i^v \in \mathcal{R}^{128}$. Similarly, we have the input $p_j^{in} \in \mathcal{R}^{128}$ ($j \in [1, l]$) for each product. Taking the micro-video sequence as an example, following the inputs of micro-videos, there is a multi-head attention block, where the Query, Key, and Value are the same. In particular, they are $V^{in} = \{v_1^{in}..., v_n^{in} | \forall v_i^{in} \in \mathcal{R}^{128}\}$. Finally, we have the representation $V^{intra} \in \mathcal{R}^{n*128}$ of a micro-video sequence, which captures the user's sequential behaviors on micro-videos. Similarly, we have $P^{intra} \in \mathcal{R}^{l*128}$ as the representation of a product sequence.

3.3 Cross-Domain Contrastive Learning

On e-commerce platforms, there are a large number of users with very sparse micro-video behaviors (*e.g.,* less than 3 watched micro-videos in the last 7 days). If we train the cross-domain recommendation model from scratch, these users' behavior data will affect the model's cross-domain knowledge transferability (c.f. the in-depth analysis in Section 4.5). Therefore, we propose a Cross-domain Contrastive Learning (CCL) algorithm to pre-train the micro-video and product multi-modal sequence encoders. The motivation is that an

Algorithm 1: Cross-Domain Contrastive Learning

Input: Collections of micro-video and product sequential behaviors V^s and P^s; dictionary D_{vs} and D_{ps}; encoders f_{vs}, f_{ps}, z_{vs}, z_{ps}; dictionary size K; learning rate γ; mini-batch size M; momentum m.

Output: Optimal solution $\theta_q^{vs}, \theta_q^{ps}, \theta_k^{vs}, \theta_k^{ps}$

1 Initialize parameters, $\theta_q^{vs}, \theta_q^{ps}, \theta_k^{vs}, \theta_k^{ps} \sim Uniform(0,1)$

2 Load a dictionary at random for D_{vs}, $D_{vs} \leftarrow \{v_1^s, ..., v_K^s\} \sim Random(V^s)$

3 Load a dictionary at random for D_{ps}, $D_{ps} \leftarrow \{p_1^s, ..., p_K^s\} \sim Random(P^s)$

4 Encode dictionary samples, $k_i^{vs} \leftarrow g(z_{vs}(v_i^s)), \forall v_i^s \in D_{vs}$, $k_i^{ps} \leftarrow g(z_{ps}(p_i^s)), \forall p_i^s \in D_{ps}$

5 **for** *each epoch* **do**

6 **for** *each mini-batch* **do**

7 Load a mini-batch of micro-video sequence, $B_{vs} \leftarrow \{v_1^s, ..., v_M^s\} \sim V^s$

8 Load a mini-batch of product sequence, $B_{ps} \leftarrow \{p_1^s, ..., p_M^s\} \sim P^s$

  `// Update dictionaries`

9 Encode mini-batch samples, $k_i^{vs} \leftarrow g(z_{vs}(v_i^s)), \forall v_i^s \in B_{vs}$

10 Encode mini-batch samples, $k_i^{ps} \leftarrow g(z_{ps}(p_i^s)), \forall p_i^s \in B_{ps}$

11 Update $D_{vs} \leftarrow ENQUEUE(DEQUEUE(D_{vs}, B_{vs}))$

12 Update $D_{ps} \leftarrow ENQUEUE(DEQUEUE(D_{ps}, B_{ps}))$

  `// Cross-domain contrastive coding`

13 Encode mini-batch samples, $q_i^{vs} \leftarrow g(f_{vs}(v_i^s)), \forall v_i^s \in B_{vs}$

14 Encode mini-batch samples, $q_i^{ps} \leftarrow g(f_{ps}(p_i^s)), \forall p_i^s \in B_{ps}$

15 Compute the posterior, $prob(y_i^{vs}|v_i^s, p_i^s, D_{vs}) \leftarrow Eq.\ 5$

16 Compute the posterior, $prob(y_i^{ps}|p_i^s, v_i^s, D_{ps}) \leftarrow Eq.\ 6$

  `// Update model parameters`

17 Update $\theta_q^{vs} \leftarrow Eq.\ 7$, $\theta_q^{ps} \leftarrow Eq.\ 8$

18 Momentum update $\theta_k^{vs}, \theta_k^{ps} \leftarrow Eq.\ 9$

active user's actions on the e-commerce platform within a session have a certain consistency of intention [14]. The objective of CCL is to encourage the representations of a micro-video sequence and a product sequence to be similar, if they come from the same session of an active user. Inspired by the success of MoCo [17], we design a momentum contrastive learning algorithm for the cross-domain recommendation scenario. As shown in Figure 2 (a), we build two queue-based dictionaries for micro-video and product sequence per mini-batch, where "keys" are randomly sampled from the data. The sequence encoders are trained to perform dictionary look-up: an encoded "query" should be similar to the value of its matching key and dissimilar to others. This training objective maximizes a lower bound of the mutual information between different domains.

Let $V^s = \{v_1^s, ..., v_L^s\}$ and $P^s = \{p_1^s, ..., p_L^s\}$ denote the collections of users' behavior sequences, where v_i^s and p_i^s are a user's sequential actions on micro-videos and products, respectively. Note that each pair (v_i^s, p_i^s) is from the same session of a user. We define query encoders f_{vs} and f_{ps} and key encoders z_{vs} and z_{ps} for micro-video and product sequence behaviors, respectively. The learnable parameters of query encoders are $\{\theta_q^{vs}, \theta_q^{ps}\}$, and the parameters of key encoders are $\{\theta_k^{vs}, \theta_k^{ps}\}$. The structure of these sequence encoders are described in the Section 3.2.3. These encoders compute representations of micro-video and product sequences as queries and keys,

$$\begin{aligned} q^{vs} &= g(f_{vs}(v^{s,query})), \quad k^{vs} = g(z_{vs}(v^{s,key})), \\ q^{ps} &= g(f_{ps}(p^{s,query})), \quad k^{ps} = g(z_{ps}(p^{s,key})), \end{aligned} \tag{4}$$

where $g(\cdot)$ is the max pooling function. Finally, we have $\{q^{vs}, k^{vs}, q^{ps}, k^{ps}\} \in \mathcal{R}^{128}$.

Algorithm 1 describes the details of the proposed CCL algorithm. The dictionaries D_{vs} and D_{ps} are initialized with K randomly drawn samples from V^s and P^s, respectively (*lines* 2–4). For each iteration within an epoch, we load new mini-batch samples and enqueue them into dictionaries (*lines* 7 – 12). Given the updated D_{vs} and D_{ps}, we perform cross-domain contrastive coding. For micro-video sequence representations to product sequence representations, we compute the posteriors of all micro-video sequence samples $v_i^s \in B_{vs}$ with respect to the negative samples in the product sequence dictionary D_{ps} as follows,

$$prob(y_i^{vs}|v_i^s, p_i^s, D_{vs}) = \frac{exp(q_i^{vs} \cdot k_i^{ps}/\tau)}{\sum_{j=1}^{K} exp(q_i^{vs} \cdot k_j^{ps}/\tau)}, \forall i \in [1, M], \tag{5}$$

where M is the mini-batch size, and τ is a temperature term. The posterior is defined over a cross-domain space with one positive and K negative pairs. Similarly, we have posteriors for product sequence representations to micro-video sequence representations (*lines* 15 – 16),

$$prob(y_i^{ps}|p_i^s, v_i^s, D_{ps}) = \frac{exp(q_i^{ps} \cdot k_i^{vs}/\tau)}{\sum_{j=1}^{K} exp(q_i^{ps} \cdot k_j^{vs}/\tau)}, \forall i \in [1, M]. \tag{6}$$

Next, we back-propagate gradients only to the query encoders f_{vs} and f_{ps} according to the cross-entropy loss (*line* 17),

$$\theta_q^{vs} \leftarrow \theta_q^{vs} - \gamma \nabla_\theta \mathcal{L}_{CE}(-\log prob(y^{vs}|\cdot), y_{gt}^{vs})|_{\theta=\theta_q^{vs}}, \tag{7}$$

$$\theta_q^{ps} \leftarrow \theta_q^{ps} - \gamma \nabla_\theta \mathcal{L}_{CE}(-\log prob(y^{ps}|\cdot), y_{gt}^{ps})|_{\theta=\theta_q^{ps}}, \tag{8}$$

where γ is the learning rate, and y_{gt}^{vs} is the ground truth label indicating whether (v_i^s, p_j^s) is a pair. After that, we apply momentum update [17] to the parameters of key encoders z_{vs} and z_{ps} (*line* 18),

$$\theta_k^{vs} \leftarrow m\theta_k^{vs} + (1-m)\theta_q^{vs}, \quad \theta_k^{ps} \leftarrow m\theta_k^{ps} + (1-m)\theta_q^{ps}, \tag{9}$$

where m is the momentum value. The momentum update allows the dictionaries to change their states slowly, thus making them consistent across iterations. In summary, we perform bi-directional contrastive coding and train the whole model end-to-end. Finally, CCL generates pre-trained sequence encoders, which can learn discriminative micro-video and product sequence representations.

3.4 Sequential Multi-Modal Information Transfer Network

The pre-trained sequence encoders bridge the domain gap at the sequence representation level by maximizing the mutual information between different domains. In our problem setting, we need to predict whether the user likes a specific micro-video (*e.g.,* target micro-video). Therefore, it requires us to further analyze the relationship between the target micro-video and the user's cross-domain behaviors. This is challenging because users' behaviors in micro-video domain are more sparse than their behaviors in the product domain. Moreover, users' behavior patterns are also largely different in these two domains. To solve these problems, we propose the Sequential Multi-Modal Information Transfer Network (SEMI). Specifically, SEMI consists of a hierarchical attention mechanism that can fully integrate and enhance users' interests in the two domains and automatically select relevant items in these domains to capture users' preferences dynamically.

As shown in Figure 2 (b), we first initialize the parameters of two sequence encoders pre-trained by the CCL algorithm, and generate intra-domain representations $V^{intra} \in \mathcal{R}^{n*128}$ and $P^{intra} \in \mathcal{R}^{l*128}$ for micro-video and product sequential behaviors, respectively. Then, we feed $[V^{intra}, P^{intra}] \in \mathcal{R}^{(n+l)*128}$ into an additional multi-head attention block to capture the inter-domain dependencies, where $[\cdot]$ denotes concatenation. The Query, Key, and Value of this inter-domain multi-head attention block are the same, that is, $[V^{intra}, P^{intra}]$. We have the user cross-domain sequential representations $\{V^{inter} \in \mathcal{R}^{n*128}, P^{inter} \in \mathcal{R}^{l*128}\}$. Then, we need to evaluate the similarity between the target micro-video and the sequences in the two domains respectively, and generate the domain-specific user representation in each domain. Thus, we utilize another multi-head attention blocks over the target micro-video and $\{V^{inter}, P^{inter}\}$, respectively. The representation of target micro-video v_t^{in} is the sum of timestamp embedding and pre-extracted multi-modal features. In these two multi-head attention blocks, Query is the target micro-video, and Key and Value are the same, that is, V^{inter}/P^{inter}. Finally, we can obtain the user representations u_v and u_p in the micro-video domain and product domain.

In the experiments, we find that concatenation is the most efficient and powerful way to combine multiple representation vectors. Therefore, the concatenation of u_v, u_p, and v_t^{in} are fed into a multi-layer perceptron to predict the probability $\hat{y}$ that the user would like to watch the target micro-video. The following cross-entropy loss is used to learn the model parameters,

$$\mathcal{L}(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log (1 - \hat{y}), \tag{10}$$

where $y \in \{0, 1\}$ is the ground-truth label indicating whether the user watches the micro-video completely or not.

4 EXPERIMENTS

To demonstrate the effectiveness of the proposed method, we perform extensive offline experiments on a large-scale dataset collected from Taobao, which is one of the world's largest e-commerce platforms. We also deploy the proposed method on Taobao platform and perform online A/B testing to demonstrate its effectiveness.

4.1 Datasets

A large-scale dataset is collected from a micro-video consumption scenario, which serves about 10 million users every day on Taobao. This dataset covers 8 consecutive days of users' interacting records on micro-videos from the Taobao mobile App's log during December 2020. Each record consists of a record ID, a user ID, a micro-video ID, a timestamp, an interaction flag (watched the micro-video completely or not), and two sequences of the user's completely watched micro-videos and clicked products within recent 7 days, respectively. Moreover, each sequence consists of multiple tuples of (pre-extracted multi-modal features for the micro-video/product, timestamp). Here, we only reserve the records with at least one watched micro-video or five clicked products. In addition, we truncate the sequence lengths of the micro-video domain and product domain into 50, respectively, if the sequence length is larger than 50. We keep all positive records (*i.e.,* the user watched the micro-video completely) and randomly sample a subset of negative records (*i.e.,* the user watched the micro-video but not completely) to keep the ratio of positive and negative records $1 : 1$ nearly. We hold the first seven days as the training set and leave the eighth day for testing. Finally, we have about 10.39×10^8 and 2.49×10^8 samples for training and test datasets, respectively. Furthermore, we select samples in training data for CCL pre-training. Specifically, we keep records in the training set where the micro-video sequence length is larger than 10. Then, we do deduplication for these records according to the user ID dimension and keep one record with the longest average length of two domains' sequences for each user. In this way, we have about 2.83×10^7 cross-domain sequence pairs with similar potential interests for CCL pre-training. More statistics of the training set, test set, and CCL pre-training set are listed in Appendix A.3.

4.2 Experimental Settings

4.2.1 Baseline Methods. We compare the proposed SEMI transfer network with CCL pre-training strategy (denoted as SEMI w. CCL) with the following baseline methods: 1) **DIN** [39]: This method uses the embedding product attention mechanism to learn adaptive representations of users' behaviors; 2) **DIEN** [38]: This method introduces an interest-evolving layer on the top of DIN to capture users' dynamic interests over time; 3) **BST** [5]: This approach leverages the Transformer structure with time information for e-commerce product recommendation; 4) **YoutubeNet** [10]: This method is proposed to recommend videos in YouTube. It obtains user representations by simply averaging the item embeddings in the user behavior sequence; 5) **BERT4SessRec** [6]: It is a video recommendation model with BERT-like structure and employs a pre-training strategy to better model users' sequential behaviors; 6) **THACIL** [8]: It is a state-of-the-art micro-video recommendation method modeling users' sequential behaviors. It leverages item- and category-attention mechanisms to capture the diverse and fine-grained interests respectively; 7) **CoNet** [1]: This method adds cross-connection units on two MLPs with shared user embeddings across domains to enable dual knowledge transfer; 8) π-**Net** [27]: It uses RNNs to encode sequences and a gating mechanism to transfer information between sequences in different domains; 9) **MiNet** [29]: This cross-domain recommendation method jointly models three

Table 1: Performances of different methods on e-commerce micro-video recommendations. * indicates the improvement over the second-best baseline is statistically significant with $p < 0.01$ using *t-test*.

Group	Method	Metrics	
		AUC	HR@3
Single-Domain	DIN	0.7027	0.4524
	DIEN	0.7067	0.4577
	BST	0.7128	0.4651
	YouTubeNet	0.6833	0.4265
	BERT4SessRec	0.7182	0.4700
	THACIL	0.7019	0.4519
Cross-Domain	CoNet	0.6749	0.4220
	π-Net	0.7241	0.4784
	MiNet	0.7244	0.4785
	Cross-BST	0.7202	0.4752
	SEMI w. CCL	**0.7396***	**0.4997***

types of user interests from different domains and utilizes two levels of attentions to fuse these interests; 10) **Cross-BST**: This method is extended from BST [5]. It is the baseline previously used in online services of Taobao. In this method, we fuse micro-videos and products as one behavior sequence, which is ordered by timestamp.

We run each method three times and compute the mean to eradicate any discrepancies. Hyper-parameter settings and implementation details of our method and baseline methods are introduced in Appendix A.1 and A.2, respectively.

4.2.2 Evaluation Metrics. We adopt two commonly used performance metrics for evaluation, *i.e.,* Area Under the receiver operating characteristic Curve (AUC) and Hit Ratio at rank K (HR@K). Specifically, we calculate HR@K according to the target micro-video and 300 randomly selected micro-videos that are not watched by the user. As the online system of Taobao requests three recommended results at a time, we set K to 3. The larger value of AUC and HR@K indicates better performances.

4.3 Overall Performance Comparison

Table 1 summarizes the e-commerce micro-video recommendation performances achieved by different methods. From Table 1, we can make the following observations.

The single-domain methods can be divided into two sub-groups: classical e-commerce recommendation methods (*i.e.,* DIN, DIEN, and BST) and video/micro-video recommendation methods (*i.e.,* YouTubeNet, BERT4SessRec, and THACIL). Comparing these two sub-groups, we can note that the classical e-commerce recommendation methods are comparable with the video/micro-video recommendation methods. This indicates that one crucial point for the micro-video recommendation task is to depict users' sequential behaviors. In particular, BST and BERT4SessRec achieve better performances. The potential reason is that the transformer architecture is effective in capturing sequence patterns. Thus, we also adopt transformer architectures in our method.

Table 2: Results of ablation studies. SEMI-Product denotes only using product domain behaviors as inputs. SEMI-Micro-Video denotes only using micro-video domain behaviors as inputs. "w. CCL" indicates training the model with CCL pre-training. "w/o CCL" indicates training the model from scratch without CCL pre-training.

Group	Method	Metrics	
		AUC	HR@3
Baselines	BERT4SessRec	0.7182	0.4700
	MiNet	0.7244	0.4785
w/o CCL	SEMI-Product w/o CCL	0.6797	0.4250
	SEMI-Micro-Video w/o CCL	0.7168	0.4694
	SEMI w/o CCL	0.7238	0.4790
w. CCL	SEMI-Product w. CCL	0.6990	0.4442
	SEMI-Micro-Video w. CCL	0.7161	0.4689
	SEMI w. CCL	**0.7396**	**0.4997**

Secondly, the cross-domain methods usually outperform single-domain methods, except for CoNet, which is a non-sequential method. This result verifies our motivation that product domain behaviors and micro-video domain behaviors are interrelated for e-commerce scenarios. The users' behaviors in the product domain can supplement and enhance the prediction of users' interests in the e-commerce micro-video domain. Moreover, it is essential to design and learn a cross-domain sequential model.

Thirdly, according to the results, it is obvious that our proposed SEMI transfer network with CCL pre-training strategy (SEMI w. CCL) outperforms all the baselines, in terms of all the evaluation metrics. Specifically, compared to the second-best baseline (*i.e.,* MiNet), SEMI w. CCL achieves 2.10% and 4.43% better performances, in terms of AUC and HR@3, respectively. These experimental results demonstrate the effectiveness of the proposed method in the e-commerce micro-video recommendation task.

4.4 Ablation Study

Moreover, we also perform ablation study to investigate the impacts of different components of the proposed method. Table 2 summarizes the results of ablation studies.

Effect of Different Domain Behaviors. For both groups in Table 2, the cross-domain method (*i.e.,* SEMI) significantly outperforms the single-domain methods with only micro-video sequence behaviors (*i.e.,* SEMI-Micro-Video) or product sequence behaviors (*i.e.,* SEMI-Product). This further demonstrates the necessity of cross-domain interest modeling for the e-commerce micro-video recommendation task. Moreover, SEMI-Micro-Video w/o CCL gains 5.46% AUC and 10.45% HR@3 against SEMI-Product w/o CCL. It confirms that there is a noticeable domain gap between the micro-video and product domains.

Effect of SEMI. To study the performances of SEMI transfer network, we train the SEMI transfer network from scratch without CCL (denoted by SEMI w/o CCL). We can observe that SEMI w/o CCL outperforms the best single-domain baseline BERT4SessRec. Compared with the best cross-domain baseline MiNet, SEMI w/o

CCL also achieves comparable results, especially in terms of HR@3. These results verify the effectiveness of SEMI transfer network in introducing cross-domain information.

Effect of CCL. Firstly, SEMI w. CCL achieves improvements as much as 2.18% for AUC and 4.32% for HR@3, compared to SEMI w/o CCL. This demonstrates the effectiveness of the CCL pre-training strategy for cross-domain interests transfer. Secondly, comparing SEMI-Product w. CCL to SEMI-Product w/o CCL, SEMI-Product w. CCL gains 2.84% AUC and 4.52% HR@3. This confirms that the CCL can help learn interests in the micro-video domain from the product domain behaviors. Finally, comparing SEMI-Micro-Video w. CCL to SEMI-Micro-Video w/o CCL, it does not achieve noticeable improvements and sometimes even exhibits worse. One possible reason is that CCL forces the micro-video sequence encoder to express some information in the product domain irrelevant to the micro-video recommendation task.

4.5 In-Depth Analysis for CCL

In this section, we analyze the impacts of different pre-training settings on the recommendation performances and present an in-depth analysis of the effectiveness of CCL.

Effect of Dictionary Size. Figure 3 shows how the dictionary size of CCL affects the micro-video recommendation performance. Firstly, when the dictionary size is less than about 2×10^4, the recommendation performance improves rapidly with the increase of the dictionary size. Then, the recommendation performance benefits a little from larger dictionaries up to a threshold (at about 5×10^4). Finally, the recommendation performance decreases when the dictionary size is larger than about 5×10^4. This observation is consistent with previous empirical findings of contrastive learning [17]. In particular, there are performance limits by increasing the dictionary size of the randomly-sampling strategy. A better way to build the contrastive learning dictionary is left for future study.

Effect of Pre-Training Dataset. We investigate the effects of the minimum micro-video sequence length of the pre-training dataset for CCL, which is denoted by MLEN in the following. Figure 4 shows that SEMI w. CCL achieves the best performance when MLEN equals 10. No matter whether MLEN is smaller or larger than 10, the performance would decline. One possible reason is that users' micro-video interests with very sparse micro-video behaviors may not be robust and convincing. If these data directly participate in CCL for pre-training or SEMI for training from scratch, the model's transfer learning ability for different domain behaviors may be hurt. For example, SEMI w. CCL even performs slightly worse than SEMI w/o CCL, when MLEN equals 1. On the contrary, for users with very dense micro-video behaviors (*e.g.,* MLEN equals 20), their micro-video domain interests are relatively dynamic, and the relevant samples are relatively less. The sequence encoders pre-trained by CCL will be biased if only utilizing these active users in the micro-video domain.

Improvements for Different User Groups. Figure 5 shows the performances of SEMI w. CCL and SEMI w/o CCL for users with different activeness in the micro-video domain. On the whole, the CCL pre-training strategy brings gains to all user groups. Especially for the inactive users in the micro-video domain, SEMI w. CCL achieves the highest gains. Take users with one or less micro-video action as an example, SEMI w. CCL achieves an average gain of 4.12%. For micro-video domain active users, the gains brought by CCL are relatively small. It is not surprising, because the micro-video domain's rich behaviors for these active users play a leading role in recommendation performances.

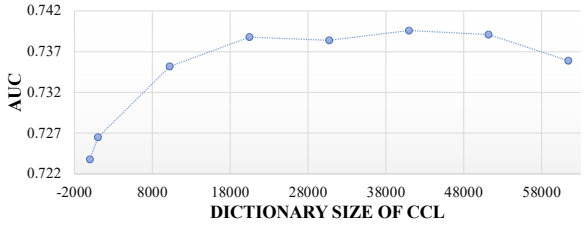


Figure 3: Recommendation performances w.r.t. different dictionary size of CCL. Particularly, when the size equals 0 (the leftmost node of the curve), it means SEMI w/o CCL.

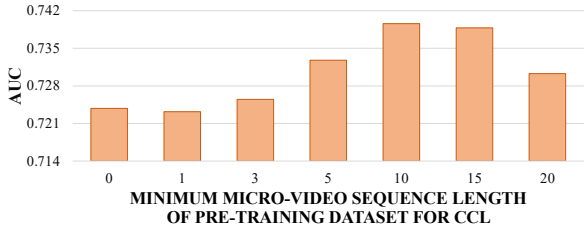


Figure 4: Recommendation performances w.r.t. minimum micro-video sequence length of pre-training dataset for CCL. Particularly, when the minimum length equals 0, it means SEMI w/o CCL.

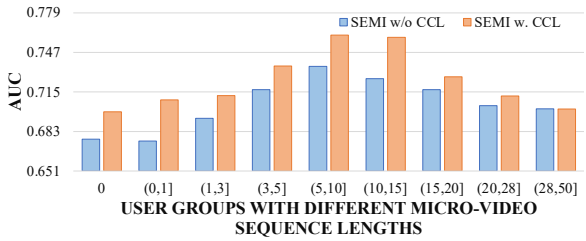


Figure 5: Recommendation performances w.r.t. user groups with different micro-video sequence lengths.

4.6 Online Experiments

We perform a careful online A/B testing on Taobao from 2020-11 to 2020-12, which is under the bucket tests. One bucket is selected for baseline and another bucket for our model. Each bucket serves about 0.5 million users per day. For the whole bucket, the goal is to increase the user stickiness and activity. During nearly two months of A/B testing, our approach contributes up to 9.32% number of browsing videos per user, 10.45% dwell time per user, and 12.10% complete-watch ratio per video, compared with the online baseline (*i.e.,* Cross-BST). These online benefits from our method are crucial for the micro-video commercial layout of Taobao. For example, advertising videos will have more broadcast opportunities with the

increased number of users watching micro-videos. The proposed method has already been fully deployed online in December 2020 and serves over 30 million users on mobile Taobao every day. The details of the online deployment are introduced in Appendix A.4.

5 CONCLUSIONS

This paper studies the e-commerce micro-video recommendation task, which is a crucial application of e-commerce platforms. As most e-commerce micro-videos are about products, we propose to utilize users' behaviors in the product domain to assist the micro-video recommendation task. Specifically, we propose a sequential multi-modal information transfer network to identify and enhance user preferences in the micro-video domain. To better bridge the gap between the two domains, we propose a cross-domain contrastive learning pre-training strategy to learn the sequence behavior encoders. Experiments on large-scale industrial dataset and long-term online A/B testing results demonstrate the effectiveness of the proposed method, compared with state-of-the-art methods. Furthermore, the proposed method has been fully deployed online on Taobao and serves over 30 millions of users every day.

6 ACKNOWLEDGMENTS

This research is supported, in part, by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore. This research is also supported, in part, by the National Natural Science Foundation of China under Grants 61836011 and 62021001.

REFERENCES

[1] Guangneng Hu andd Yu Zhng and Qiang Yang. 2018. CoNet: Collaborative Cross Networks for Cross-Domain Recommendation. In *CIKM*. ACM, 667–676.
[2] Shumeet Baluja, Rohan Seth, Dharshi Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. 2008. Video Suggestion and Discovery for Youtube: Taking Random Walks Through the View Graph. In *WWW*. ACM, 895–904.
[3] Bisheng Chen, Jingdong Wang, Qinghua Huang, and Tao Mei. 2012. Personalized Video Recommendation through Tripartite Graph Propagation. In *MM*. ACM, 1133–1136.
[4] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*. ACM, 335–344.
[5] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. Behavior Sequence Transformer for E-Commerce Recommendation in Alibaba. In *DLP-KDD*. ACM, 1–4.
[6] Xusong Chen, Dong Liu, Chenyi Lei, Rui Li, Zheng-Jun Zha, and Zhiwei Xiong. 2019. BERT4SessRec: Content-Based Video Relevance Prediction with Bidirectional Encoder Representations from Transformer. In *MM*. ACM, 2597–2601.
[7] Xusong Chen, Dong Liu, Zhiwei Xiong, and Zheng jun Zha. 2020. Learning and Fusing Multiple User Interest Representations for Micro-Video and Movie Recommendations. In *Transactions on Multimedia*, Vol. 23. IEEE, 484–496.
[8] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhouand Zhiwei Xiong, and Yan Li. 2018. Temporal Hierarchical Attention at Category-and Item-level for Micro-video Click-through Prediction. In *MM*. ACM, 1146–1153.
[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. arXiv:1909.11740
[10] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for Youtube Recommendations. In *RecSys*. ACM, 191–198.
[11] Peng Cui, Zhiyu Wang, and Zhou Su. 2014. What Videos are Similar with You?: Learning a Common Attributed Representation for Video Recommendations. In *MM*. ACM, 597–606.
[12] Qinxu Ding, Yong Liu, Chunyan Miao, Fei Cheng, and Haihong Tang. 2021. A Hybrid Bandit Framework for Diversified Recommendation. In *AAAI*. 4036–4044.
[13] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems. In *WWW*. ACM, 278–288.
[14] Yufei Feng, Fuyu Lv, Weichen Shen, Menghan Wang, Fei Sun, Yu Zhu, and Keping Yang. 2019. Deep Session Interest Network for Click-through Rate Prediction. In *IJCAI*. 2301–2307.
[15] Liang Hu, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu, and Can Zhu. 2013. Personalized Recommendation via Cross-domain Triadic Factorization. In *WWW*. ACM, 596–606.
[16] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What Aspect Do You Like: Multi-scale Time-aware User Interest Modeling for Micro-video Recommendation. In *MM*. ACM, 3487–3495.
[17] Yuxin Wu Saining Xie Ross Girshick Kaiming He, Haoqi Fan. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*. IEEE.
[18] Chenyi Lei, Shouling Ji, and Zhao Li. 2019. TiSSA: A Time Slice Self-Attention Approach for Modeling Sequential User Behaviors. In *WWW*. ACM, 2964–2970.
[19] Yongq Li, Meng Liu, Jianhua Yin, Chaoran Cui, Xin-Shun Xu, and Liqiang Nie. 2019. Routing Micro-videos via A Temporal Graph-guided Recommendation System. In *MM*. ACM, 384–400.
[20] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. 2019. A Pareto-Efficient Algorithm for Multiple Objective Optimization in E-Commerce Recommendation. In *RecSys*. ACM, 20–28.
[21] Shang Liu and Zhenzhong Chen. 2019. Sequential Behavior Modeling for Next Micro-Video Recommendation with Collaborative Transformer. In *ICME*. IEEE, 460–465.
[22] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-Video Co-Attention Network for Personalized Micro-video Recommendation. In *WWW*. ACM, 3020–3026.
[23] Yong Liu, Yingtai Xiao, Qiong Wu, Chunyan Miao, Juyong Zhang, Binqiang Zhao, and Haihong Tang. 2020. Diversified interactive recommendation with implicit feedback. In *AAAI*. 4932–4939.
[24] Babak Loni, Yue Shi, Martha Larson, and Alan Hanjali. 2014. Cross-Domain Collaborative Filtering with Factorization Machines. In *CERI*. 456–661.
[25] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. arXiv:2002.06353
[26] Jingwei Ma, Jiahui Wen, Mingyang Zhong, Weitong Chen, Xiaofang Zhou, and Jadwiga Indulska. 2019. Multi-Source Multi-Net Micro-Video Recommendation with Hidden Item Category Discovery. In *DASFAA*. Springer, 1464–1472.
[27] Muyang Ma, Pengjie Ren, Yujie Lin, Zhumin Chen, Jun Ma, and Maarten de Rijke. 2019. pi-Net: A Parallel Information-sharing Network for Shared-account Cross-domain Sequential Recommendations. In *SIGIR*. ACM, 685–694.
[28] Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. 2017. Cross-Domain Recommendation: An Embedding and Mapping Approach. In *IJCAI*. 2464–2470.
[29] Wentao Ouyang, Xiuwu Zhang, Lei Zhao, Jinmei Luo, Yu Zhang, Heng Zou, Zhaojie Liu, and Yanlong Du. 2020. MiNet: Mixed Interest Network for Cross-Domain Click-Through Rate Prediction. In *CIKM*. ACM, 2669–2676.
[30] Guorui Zhou Xiaoqiang Zhu Kun Gai Qi Pi, Weijie Bian. 2019. Practice on Long Sequential User Behavior Modeling for Click-Through Rate Prediction. In *KDD*. ACM, 2671–2679.
[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *NIPS*. ACM, 6000–6010.
[32] Peng Wang, Yunsheng Jiang, Chunxu Xu, and Xiaohui Xie. 2019. Overview of Content-Based Click-Through Rate Prediction Challenge for Video Recommendation. In *MM*. ACM, 2593–2596.
[33] Shichao Wang, Danyang Song, Xi Chen, Tianqi Shi, and Haihong Tang. 2020. What Message an E-commerce Video Should and Should not Communicate? The Influence of Video Content on Its Business Performance. In *The Workshop on e-Business*.
[34] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *MM*. ACM, 1437–1445.
[35] YanxiangHuang, BinCui, JieJiang, KunqianHong, WenyuZhang, and YiranXie. 2016. Real-Time Video Recommendation Exploration. In *SIGMOD*. ACM, 35–46.
[36] Yinan Zhang, Yong Liu, Peng Han, Chunyan Miao, Lizhen Cui, Baoli Li, and Haihong Tang. 2020. Learning personalized itemset mapping for cross-domain recommendation. In *IJCAI*. 2561–2567.
[37] Cheng Zhao, Chenliang Li, Rong Xiao, Hongbo Deng, and Aixin Sun. 2020. CATN: Cross-Domain Recommendation for Cold-Start Users via Aspect Transfer Network. In *SIGIR*. ACM, 229–238.
[38] 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 *AAAI*. 5941–5948.
[39] 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 *KDD*. ACM, 1059–1068.

A APPENDIX

In the appendix, we first present the implementation details of the proposed method and baseline methods. Then, we introduce more details about the large-scale dataset collected from Taobao. After that, we introduce our online deployment system and pipeline. Finally, we provide some details of the multi-modal feature extractors used in this work.

A.1 Implementation of the Proposed Method

We set the number of heads and dimensions of feed-forward networks in all multi-head attention blocks to 8 and 128, respectively. The Multi-Layer Perceptron (MLP) used in SEMI is set to $128 \rightarrow 1$. For both SEMI and CCL, the batch size is set to 1024, and the optimizer is Adam. For CCL, we set initial learning rate $\gamma = 10^{-3}$, dictionary size $K = 40 \times 1024$, momentum $m = 0.999$, and temperature $\tau = 0.7$. For SEMI, we set the initial learning rate $\gamma = 10^{-3}$ and adopt a dropout with a rate 0.1 after the final MLP. The proposed method is implemented with TensorFlow 1.4 in Python 2.7, and is conducted on Alibaba's distributed cloud platform[1] which contains thousands of workers. Every two workers share an NVIDIA Tesla P100 GPU with 16GB of memory.

A.2 Implementation of Baseline Methods

We implement each baseline method as follows and report its performances under its optimal settings.

- **DIN [39], DIEN [38]:** We implement DIN and DIEN following the original papers and public code$_1$[2] and code$_2$[3]. We use the same multi-modal embeddings with our method as inputs.
- **BST [5], Cross-BST:** We implement BST according to the original paper. We set the multi-head attention blocks in BST as the same as SEMI, and set the number of blocks to 1. The difference of Cross-BST compared to BST is the cross-domain inputs.
- **YoutubeNet [10]:** We implement this model in our code following the original paper.
- **BERT4SessRec [6]:** We implement this method according to the original paper. We set the number of multi-head attention blocks in BERT4SessRec the same as that used in SEMI, and set the number of blocks to 2. In the pre-training stage, the mask percentage is set to 0.1.
- **THACIL [8]:** We implement this method according to original papers and the public code[4].
- **CoNet [1]:** We implement this method following the original paper and utilize the average pooling of different domains' actions as the user embedding.
- **π-Net [27], MiNet [29]:** We implement these two baselines following the original papers and public code$_1$[5] and code$_2$[6]. As these two methods need the loss from the product domain, we utilize the latest clicked product in the sequence as the training target for the product domain in the training stage.

[1]https://data.aliyun.com
[2]https://github.com/zhougr1993/DeepInterestNetwork
[3]https://github.com/YafeiWu/DIEN
[4]https://github.com/Ocxs/THACIL
[5]hhttps://github.com/mamuyang/PINet
[6]https://github.com/oywtece/minet

Table 3: Statistics of the dataset for SEMI.

Index	Dataset	
	Training	Test
#Unique Instances	10.39×10^8	2.49×10^8
#Unique Users	4.42×10^7	1.08×10^7
#Unique Target Micro-Videos	5.30×10^6	2.33×10^6
#Unique Micro-Videos in Sequences	1.14×10^7	5.20×10^6
#Unique Products in Sequences	1.12×10^8	4.88×10^7

Table 4: Statistics of the micro-video and product sequences in the dataset for SEMI.

Number	Training		Test	
	Micro-Video	Product	Micro-Video	Product
Avg. Length	26.84	44.97	27.28	44.57
Avg. #Category	16.33	19.31	16.78	19.12
Diversity	0.6084	0.4294	0.6154	0.4291
Avg. #Overlap	1.93		1.96	

Table 5: Statistics of the dataset for CCL.

Number	Min. Len. = (1, 5)		Min. Len. = (10, 5)		Min. Len. = (20, 5)	
	Micro-Video	Product	Micro-Video	Product	Micro-Video	Product
Avg. Length	23.06	44.96	31.19	46.03	35.38	46.30
Avg. #Category	13.67	18.60	19.31	19.38	22.36	19.68
Diversity	0.5924	0.4137	0.6193	0.4211	0.6319	0.4251
Avg. #Overlap	2.02		2.41		2.47	
#Instances	4.64×10^7		2.83×10^7		1.64×10^7	

A.3 Details of Experimental Datasets

This section introduces the detailed statistics of the large-scale dataset collected from Taobao. The categories of micro-videos are predicted by a content-based category-prediction model of Taobao. In the following tables, "Avg. Length" denotes the average sequence length. "Avg. #Category" denotes the average number of categories per sequence. "Diversity" denotes the diversity of the sequence and equals to "Avg. #Category"/"Avg. Length". "Avg. #Overlap" denotes the average number of overlapped categories between two corresponding sequences. "#Instances" denotes the number of training samples. "Min. Len." denotes the minimum sequence lengths of micro-videos and products (c.f. Section 4.1).

Table 3 shows the statistics of the datasets for SEMI. The huge scale of the dataset also indicates the challenge of our task. As we utilize multi-modal features in our method, our method can cover the problem that many videos in the test set have not appeared in the training set. Table. 4 further shows the statistics of the micro-video and product sequences in the dataset for SEMI. We have three observations from Table. 4. First, users on Taobao are more active in the product domain comparing with the micro-video domain, according to the Avg. Length. Second, users' interests are more dynamic in the micro-video domain than the product domain according to Diversity. Third, user behaviors in different domains may reflect similar preferences according to Avg. #OverLap. However, the values of Avg. #OverLap are very small compared to the values of Avg. #Category, which also indicates that the user behavior patterns are quite different across the two domains.

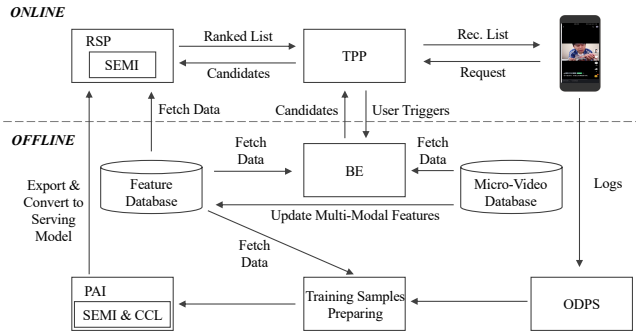


Figure 6: The deployment pipeline of SEMI w. CCL in Taobao APP.

Table. 5 shows the statistics of the datasets for CCL based on different minimum micro-video sequence length (c.f. Section 4.1). Note that different data selection rules for CCL will impact the micro-video recommendation performances (c.f. Fig. 4).

A.4 Deployment System and Pipeline

The proposed method has been fully deployed online in Taobao and serves over 30 million users every day. In this section, we introduce the online deployment pipeline of the proposed method. Fig. 6 illustrates the related system and pipeline. For the online sub-system, the Taobao Personality Platform (TPP) and Ranking Service Platform (RSP) are the main components. A typical workflow is illustrated as follows:

- When a user launches Mobile Taobao App, TPP extracts the user's latest information, including the latest clicked products, and watched micro-videos. Meanwhile, TPP utilizes the user's information to retrieve a candidate set of micro-videos from Basic Engine (BE), which is an effective recall engine in Taobao, according to several recall approaches such as collaborative filtering and embedding similarity. These candidates are fed to RSP. RSP fetches related pre-extracted multi-modal features of candidates from Feature Database and ranks the candidate set of micro-videos with a trained SEMI transfer network and returns the ranked results to TPP. Particularly, under about 2,500 service requests per second, SEMI on RSP can make inferences for about 500 candidates in batches with a delay of about 30 ms.
- During their visits to Taobao, users' behaviors are collected and saved as log data for the offline sub-system on Open Data Processing Service (ODPS) distributed platform.

The workflow and sub-modules of the offline sub-system are described as follows:

- The log instances contain the user's historical sequential clicked products and watched micro-videos with timestamps, as well as the observed micro-video currently. To date, about 2.3 billion log instances are generated by about 34 million users every day. All of these logs are saved on ODPS. Based on these logs and the ODPS platform, we prepare the training dataset for our method (c.f. Section 4.1). Moreover, the related multi-modal features are also fetched from the Feature Database.
- To implement the proposed e-commerce micro-video recommendation algorithms, we train our models on the Platform of Artificial Intelligence (PAI), which is a distributed machine learning platform in Taobao, with thousands of NVIDIA Tesla P100. For better fitting the changes in the data distribution, we re-train the models every week by the automatic dispatch system of PAI.
- To reduce the computational complexity on RSP, we pre-extract and update multi-modal features of all micro-videos and products for Feature Database. By doing this, the SEMI transfer network on RSP would avoid time cost on feature extractor during online inference.

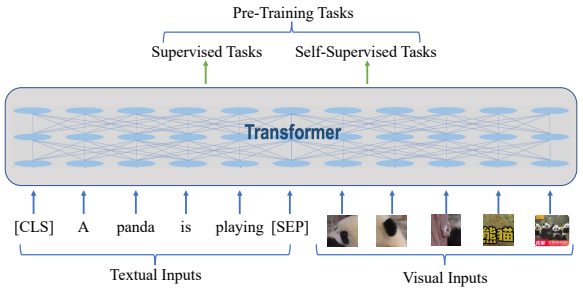


Figure 7: The framework of the general multi-modal feature extractor.

A.5 Feature Extractors

As described in Section 3.2.1, we adopt two pre-trained multi-modal feature fusing models to extract features for micro-videos and products. Fig. 7 shows the general multi-modal feature extractor framework [9, 25]. Generally, the inputs include textual and visual information. The main network structures are multi-layer transformers. The pre-training tasks include supervised tasks (*e.g.,* tag classifications) and self-supervised tasks (*e.g.,* visual-textual information matching). The pre-training tasks can guide the model to extract and fuse the corresponding multi-modal information. For example, the task of tag classifications will force the model to focus on those multi-modal signals related to tags. To date, this kind of framework has become the state-of-the-art of multi-modal information fusing and the standard tool in Taobao.

For micro-videos, the inputs of the pre-trained UniVL [25] model are micro-video titles and frames. Pre-training tasks include masked language modeling, masked frame modeling, video-text alignment, and tag classifications. These tasks will guide the model to learn the semantic information of micro-video frames, titles, and tags. Finally, the output of the model is a 128-dimension embedding. We pre-train UniVL with about 100 million e-commerce micro-videos.

For products, the inputs of the pre-trained UNITER [9] model are regions of cover pictures and titles. Pre-training tasks include masked language modeling, masked region modeling, image-text matching, and category classifications. These tasks will guide the model to learn the semantic information of product cover pictures, titles, and categories. Finally, the output of the model is a 128-dimension embedding. We pre-train UNITER with about 100 million e-commerce products.