This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3054525, IEEE Transactions on Multimedia

E-Commerce Storytelling Recommendation Using Attentional Domain-Transfer Network and Adversarial Pre-Training

Xusong Chen, Chenyi Lei, Dong Liu, *Senior Member, IEEE*, Guoxin Wang, Haihong Tang,
Zheng-Jun Zha, *Member, IEEE*, and Houqiang Li, *Fellow, IEEE*

***Abstract*—In e-commerce platforms, there is an emerging type of content that tells a "story" about some merchandise in the form of multimedia (text, images, video), which is named *storytelling*. Well told stories, like advertisements, can inspire users to purchase the related products. Thus, e-commerce service provider is keen to disseminate storytelling items to potentially interested users. We address this requirement by a cross-domain personalized recommendation approach. Because storytelling is a new type of content, its related user actions are much less, more sparse than product-related user actions, thus we propose to use product-domain user actions to assist the identification of user preferences and to make storytelling recommendations. Our method has two technical contributions. First, since the user behavior patterns are different across the storytelling domain and the product domain, we propose an attentional domain-transfer network, which effectively selects the relevant items in the two domains to characterize user preferences. Second, although storytelling is about product, between the two domains there is a large gap: product description is objective and categorical, like "keywords," but storytelling is close to human language. To bridge the domain gap, we propose a dual-domain contrastive adversarial learning method to pre-train the feature extractors for storytelling and product simultaneously. We conduct experiments on two industrial datasets, and the results demonstrate the advantage of our proposed method that consistently outperforms the state-of-the-art methods. Besides, our method can be used to recommend storytelling to products, which is a desired functionality for product providers. Our code and models are publicly available.**

***Index Terms*—Adversarial learning, Attentional domain-transfer network, Cross-domain recommendation, E-commerce, Storytelling recommendation**

I. Introduction

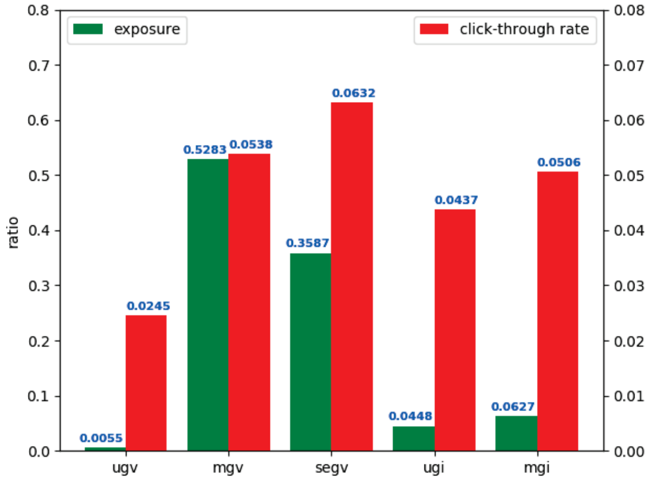


Fig. 1. Comparison between storytelling items of different sources: ugv (user-generated video), mgv (merchant-generated video), segv (shopping expert-generated video), ugi (user-generated image), and mgi (merchant-generated image). The left coordinate corresponds to exposure ratio, i.e. percentage of each category. The right coordinate corresponds to click-through rate, which is calculated by dividing the number of clicked storytellings by the number of exposed.

Storytelling is an emerging type of content in e-commerce platforms, which tells a "story" about some merchandise in the form of multimedia including text, images, and videos. As shown in Fig. 1, there is a wide spectrum of sources of e-commerce storytelling, ranging from advertisements crafted by product providers or merchants, analyses and syntheses provided by opinion leaders (also known as shopping experts), to essays contributed by grass-root users talking about e.g. their feeling about a product. For e-commerce platforms, storytelling has two important advantages. First, providing users with high-quality content enables a different experience of pan-entertainment, which greatly increases user stickiness to the platform. Second, well told stories would interest users in the related merchandise and inspire their shopping behavior. Therefore, e-commerce service provider is keen to promote the popularity of storytelling. To date, in Taobao[1] there are more than 100 million storytelling items, and new items are generated from different sources at the speed of hundreds of thousands per day. More than 50 million users have interacted with storytelling, and the daily usage time is more than 10 minutes, which leads to an estimated gross merchandise volume of millions US dollar. Accordingly, how to disseminate storytelling content to potentially interested users, especially via personalized recommendations of storytelling items, is a key requirement.

As a new type of content, storytelling is yet to reach most

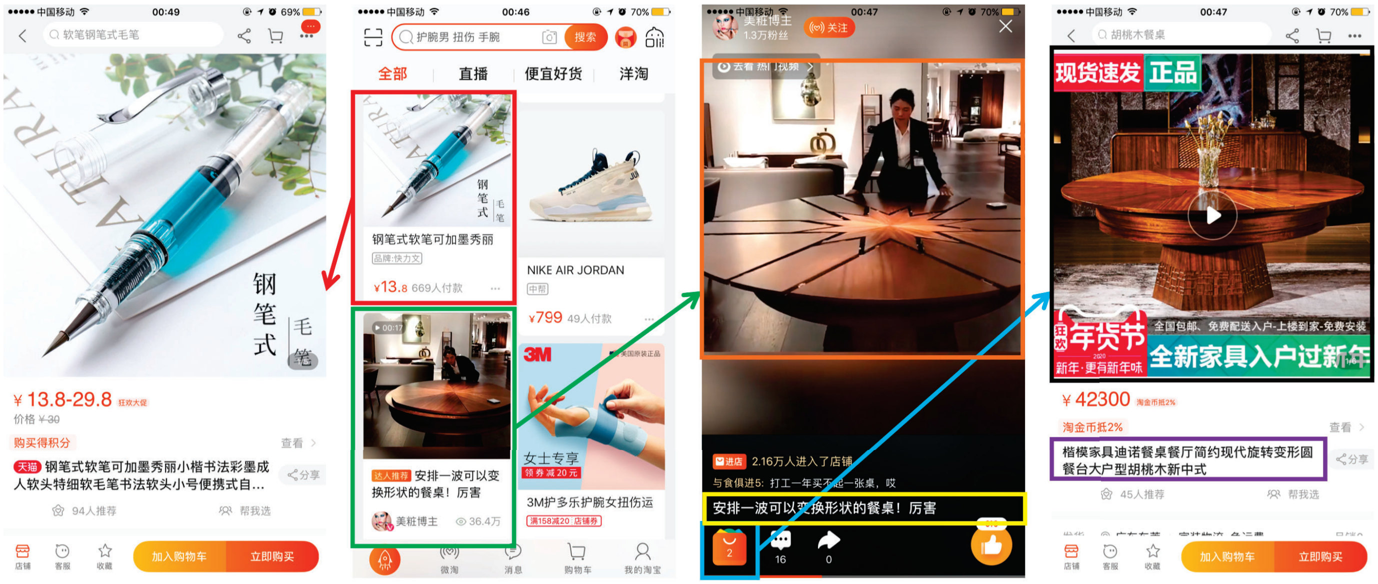
Date of current version January 21, 2021. This work was supported by the Natural Science Foundation of China under Grant 61772483.

X. Chen, C. Lei, D. Liu, Z.-J. Zha, and H. Li are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: cxs2016@mail.ustc.edu.cn; leichy@mail.ustc.edu.cn; dongeliu@ustc.edu.cn; zhazj@ustc.edu.cn; lihq@ustc.edu.cn).

C. Lei, G. Wang, and H. Tang are with Alibaba Group, Hangzhou 311121, China (e-mail: chenyi.lcy@alibaba-inc.com; xiaogong.wgx@taobao.com; piaoxue@taobao.com).

[1] www.taobao.com, one of the largest e-commerce platforms in the world.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3054525, IEEE Transactions on Multimedia



¥ 13.8-29.8 狂欢大促

价格 ¥30

购买得积分

天猫 钢笔式软笔可加墨秀丽小楷书法彩墨成人软头特细软毛笔书法软头小号便携式自...

94人推荐

钢笔式软笔可加墨秀丽

品牌:快力文

¥13.8 669人付款

NIKE AIR JORDAN

¥799 49人付款

达人推荐 安排一波可以变换形状的餐桌！厉害

美粧博主 36.4万

3M护多乐护腕女扭伤运

满158减20 店铺券

2.16万人进入了店铺

与食俱进5：打工一年买不起一张桌，哎

安排一波可以变换形状的餐桌！厉害

¥ 42300 淘金币抵2%

楷模家具迪诺餐桌餐厅简约现代旋转变形圆餐台大户型胡桃木新中式

45人推荐

(1) The content of product (2) The homepage of user browsing (3) The content of storytelling (4) The product mounted in storytelling

Fig. 2. Screenshots of the Taobao mobile app. In (2), there are both product items and storytelling items. The two kinds of items can be distinguished visually: product items have price (e.g. "¥13.8") while storytelling items have not. Clicked on the red box, we have (1). Clicked on the green box, we have (3). In (3), there is a picture on top (orange box) showing a lady operating a table, and the title (yellow box) says: "Settling down a deformable dining-table, amazing!" There is a button under the title (blue box), on which we clicked to have (4). The product in (4) is said to be *mounted* on the storytelling in (3). (3) and (4) are named a *pair*. The title of the mounted product (purple box) reads: "Kai-Mo (a trademark) furniture, Di-Nuo (a brand) dining-table, dining-room, simple, modern, rotatable, deformable, round dining-table, (suitable for) big house, walnut (material), new Chinese style." The product picture (black box) shows the dining-table as well as some discount information. Obviously, the pictures and titles of (3) and (4) are quite different, showing the gap between the storytelling domain and the product domain.

of the users, so its related user actions are few and sparse, resulting in the difficulty of "cold-start" for the personalized recommendation. Fortunately, in e-commerce platforms there are much more user actions about products, which can be used to identify user preferences and requirements. Since storytelling is also about products, the product-related user actions can be a specially important data source for storytelling recommendation. Thus, we propose to jointly use storytelling-related and product-related user actions for personalization.

When introducing user actions about products into storytelling recommendation, there are new technical challenges. First, users' intentions are different: when interacting with storytelling, user is usually expected to be "inspired" and to have fun; when interacting with product, user is more prepared to buy a specific kind of products. Accordingly, user behavior patterns are different across the storytelling domain and the product domain (see Section II-C for further analysis). Second, although storytelling is about product, their forms are quite different. Storytelling item is unstructured and can be combination of text, images, videos, etc., but product item is well structured with clearly defined attributes. Even if we consider their common properties, e.g. title and cover picture, there is still large difference. A concrete example is shown in Fig. 2, where we can compare the title and cover picture of a storytelling item (3) with those of the related (a.k.a. *mounted*[2]) product item (4). The storytelling title is close to natural language but the product title is a concatenation of keywords (which is common because merchants try to do search engine optimization). The cover picture of storytelling is captured by a normal user so it appears with a real-life context. But the cover picture of product is professionally produced to be clean and concentrated (i.e. only contains the product). In summary, there is a large gap between the storytelling domain and the product domain.

To address the challenges, we focus our study on the domain transfer and domain adaptation techniques, so as to effectively use the product-domain user actions to make storytelling-domain recommendations. In the literature, there have been several cross-domain recommendation methods [1]–[9], but they are either not designed for multimedia content, or suffering from the *item cold-start* problem. Besides, the existing multimedia recommendation methods [10]–[15] are not directly applicable to cross-domain scenarios.Thus, we propose a new recommendation method termed **A**ttentional **D**omain-transfer network with **A**dversarial **P**re-**T**raining (ADAPT) for our task. Our proposed ADAPT has two key components: **DU**al-domain contrastive **A**dversarial **L**earning (DUAL) model and **A**ttentional **D**omain-tranf**E**r **N**etwork (ADEN). Specifically, ADEN learns user interest in cross-domain to overcome the diversity of cross-domain sequential behavior. Along with the success of attention mechanisms [16], [17], we leverage the multi-head attention [16] as sequence encoder and then use target storytelling as query to dynamically select the relevant items in two domains to represent user preferences. Then, user interest representations can be aggregated by a

[2] Most of storytellings relate to one product. There are a few exceptional cases that a storytelling does not relate to any product, or a storytelling relates to multiple products.

multi-layer perceptron to predict the click-through rate of storytelling. Since ADEN ignores the domain gap between storytelling and product, we further propose DUAL model to learn the multi-modal feature representations (i.e., visual and textual feature) in order to bridge the domain gap between storytelling and product by pre-training our feature extractors. We impose dual constraints on the feature representation in order to minimize the gap among the representations of storytelling and product in the same pairs and maximize the distance in the different pairs. Moreover, we propose two domain classifiers which act as discriminators in the generative adversarial net paradigm [18] so that the feature extractors (generators) generate the domain-invariant feature for alleviating the domain gap. Finally, the pre-trained feature extractors in the DUAL model are used not only for fine-tuning click-through rate prediction tasks but also for recommending suitable storytelling to products in the e-commerce platform. For our research, we build two datasets from the real data of Taobao. Our extensive experiments verify that our proposed method is more effective than previous state-of-the-art methods.

In summary, the main contributions of this work are as follows.

- To the best of our knowledge, we are the first to investigate e-commerce storytelling recommendation, whose importance and business value have been discussed before.
- We propose to use product-domain user actions to enhance the ability of storytelling-domain recommendation. To that end, we propose an **A**ttentional **D**omain-transf**E**r **N**etwork (ADEN), which effectively selects the relevant items in the two domains to characterize user preferences.
- To bridge the gap between the storytelling domain and the product domain, we further propose a **DU**al-domain contrastive **A**dversarial **L**earning (DUAL) method to pre-train the feature extractors for storytelling and product simultaneously. DUAL-based pre-training greatly improves the ability of ADEN. Besides, DUAL can be used to recommend/match storytelling to product.
- We conduct experiments on the two datasets, including comparisons with the state-of-the-art methods as well as a comprehensive ablation study. Our code and models are publicly available[3].

The remaining sections are organized as follows. Section II presents problem setting, details of collecting dataset, and our observations from data. Section III details proposed methods. Experimental setting and results are reported in Section IV. Section V describes related work, followed by conclusions in section VI.

II. Problem, Data, and Analyses

A. Problem Setting

We consider a personalized storytelling recommendation problem, i.e. which storytelling items shall be recommended to a given user? This problem is converted to predicting the click-through probability of a user on a his/her unseen item. Since user interest is dynamic, we use his/her recent actions to help predict. As mentioned before, the recent actions include both storytelling- and product-related actions. The actions we considered are merely *clicking* in this paper. We omit the other personal data of users, like demographics, for simplicity.

Concretely, our problem is given a user and a storytelling item, where the user is represented by his/her recently clicked storytellings and products, we want to predict the probability that the user will click on the item.

B. Datasets

According to the problem setting, we build two datasets from the real data of Taobao. The dataset building strategy is the same for the two datasets, and the only difference is their raw data are at different periods. The dataset SRD-1 comes from the raw data of Nov. 27–28, 2019. The dataset SRD-2 is of Dec. 30–31, 2019, which is a special time (year end), so it is much larger than SRD-1. Our intention in building two datasets is to verify the generalization ability of the proposed method. We sample one day data for testing because our online model is updated every day.

Take SRD-1 as an example, we first randomly sample 266,666 users and corresponding storytelling-domain interactions that occurred in the Nov. 27, 2019 for training, where each interaction consists of: a user ID, a storytelling ID, a timestamp, an interaction flag (clicked or not clicked), and two sequences of the user's recently viewed storytellings and products, respectively. Each sequence further consists of multiple tuples of (storytelling or product ID, timestamp). Here we truncate the sequence length of storytelling-domain and product-domain into 20 respectively if the sequence length is larger than 20. Due to the *item cold-start* problem, the title and cover picture of each storytelling and product are also provided. Since the click behavior is sparse, we keep all positive actions (i.e., user clicked actions) and randomly keep a subset of negative actions (i.e., user viewed but not clicked actions) to keep the ratio of positive and negative actions almost 1:1. For testing, we randomly select 73,535 users (about a quarter of the number of users in training set), and all corresponding actions of these users occurred in Nov. 28, 2019 is collected to evaluate the click-through rate prediction. The procedure of collecting SRD-2 is similar with SRD-1 but in different periods.

In addition, for each storytelling we have its mounted product ID. We name a storytelling and its mounted product a *pair*[4]. The statistical nature of our datasets is shown in Table I.

C. Analyses

As mentioned before, there is a gap between the storytelling domain and the product domain. Also, user behavior patterns are different across the two domains. We conduct quantitative analyses to demonstrate.

[3] https://github.com/Ocxs/ADAPT.

[4] If a storytelling relates to multiple products, each product and the storytelling form a pair (see Section III-B).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3054525, IEEE Transactions on Multimedia

TABLE I
STATISTICS OF THE TWO DATASETS USED IN THIS WORK

Index	SRD-1		SRD-2	
	Training	Test	Training	Test
#Instances	5,083,833	4,776,319	23,321,211	19,174,368
#Users	266,666	73,535	1,230,454	292,131
#Sequences	3,796,955	1,218,539	17,088,337	4,750,474
#Storytellings	495,280	375,164	805,436	504,594
#Products	6,496,553	2,894,920	15,285,378	6,626,009

Shown are the numbers of *unique* users, sequences, storytellings, and products. Storytellings include those in instances and those in sequences. Products include those in sequences and those mounted on storytellings.

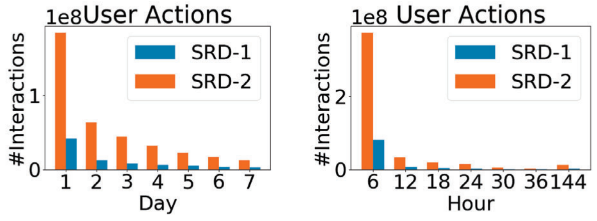
TABLE II
TITLE SIMILARITY DISTRIBUTIONS

Title similarity	SRD-1		SRD-2	
	Number	Ratio	Number	Ratio
$[0.0, 0.2)$	435,997	41.8%	368,106	29.9%
$[0.2, 0.4)$	191,482	18.4%	227,038	18.5%
$[0.4, 0.6)$	114,298	11.0%	157,221	12.8%
$[0.6, 0.8)$	114,403	11.0%	178,617	14.5%
$[0.8, 1.0]$	185,693	17.8%	298,509	24.3%

a) On the domain gap: Storytelling is usually created about the product to promote sales for merchants in the e-commerce platform. As shown in Fig. 2, we observe that although the storytelling is about the product, product title is categorical, like "keywords" but storytelling is close to human language. The reason is that users usually use keywords to search the related product on the e-commerce platform if they purchase products with clear intention. However, the storytelling is more attractive than the product if users view items without clear intention as natural language is more interesting. Moreover, we evaluate the difference of the titles between the storytellings and the products statistically. For a pair of a storytelling and its mounted product, $\mathbb{W}^s$ (resp. $\mathbb{W}^p$) is the set of words in the title of the storytelling (resp. the product). We define the metric–title similarity–by computing the number of overlapped words divided by the number of words in the shorter title between $\mathbb{W}^s$ and $\mathbb{W}^p$:

$$\text{title similarity} = \frac{|\mathbb{W}^s \cap \mathbb{W}^p|}{\min(|\mathbb{W}^s|, |\mathbb{W}^p|)} \tag{1}$$

Then we calculate the histogram of the title similarity for all the pairs in our datasets, as shown in Table II. We observe that in SRD-1 (resp. SRD-2) there are 71.2% (resp. 62.2%) pairs whose title similarity is lower than 0.6. Similar results can be inferred if using the standard Jaccard similarity. In summary, the title similarity distribution reveals a large domain gap between product and storytelling.



(a) Storytelling-related actions (b) Product-related actions

Fig. 3. Distribution of the number of user actions.

TABLE III
STATISTICS OF THE STORYTELLING AND PRODUCT SEQUENCES. SHOWN NUMBERS ARE THE AVERAGE SEQUENCE LENGTH, THE AVERAGE NUMBER OF CATEGORIES PER SEQUENCE, AND THE AVERAGE NUMBER OF OVERLAPPED CATEGORIES BETWEEN TWO CORRESPONDING SEQUENCES.

Number	SRD-1		SRD-2	
	Storytelling	Product	Storytelling	Product
Avg. Length	16.10	19.88	16.16	19.90
Avg. #Category	8.17	6.83	8.31	7.03
Avg. #Overlap	2.76		2.77	

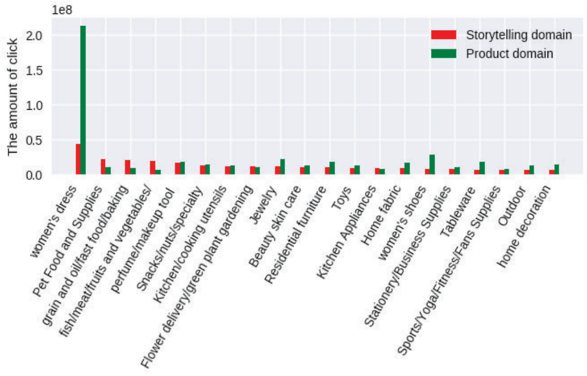


Fig. 4. Distribution of the merchandise categories in the storytelling domain and the product domain in the SRD-2 dataset.

b) On the cross-domain user behavior: We analyze on the following two aspects. Firstly, we analyze the temporal distribution of user actions. For each instance in the datasets, there are two sequences in the storytelling domain and the product domain, respectively. We calculate the time gap between the instance's timestamp and each item's timestamp in the two sequences, and illustrate the histogram of the time gap in Fig. 3. It is worth noting that the sequence length is restricted to be at most 20. We observe that the recent 20 actions in the product domain almost all occurred in the last one day, which indicates that product-domain user behaviors mostly reflect the short-term interests. The recent 20 historical actions in the storytelling domain can cover the recent seven days, which means that the user's interactions on storytellings are more sparse than on products. Secondly, we evaluate the similarity of the user's preferences between the product domain and the storytelling domain. Here we choose the overlapped categories to show because they are easy to quantify. For each product we have its category. For each storytelling, we use the category of its mounted product. Thus we can compare the category distribution in a user's product-domain sequence and storytelling-domain sequence from local and global perspectives. From a local perspective, we calculate the number of overlapped categories for each user. The statistical results are shown in Table III. We can observe that there exist overlapped categories between product sequence

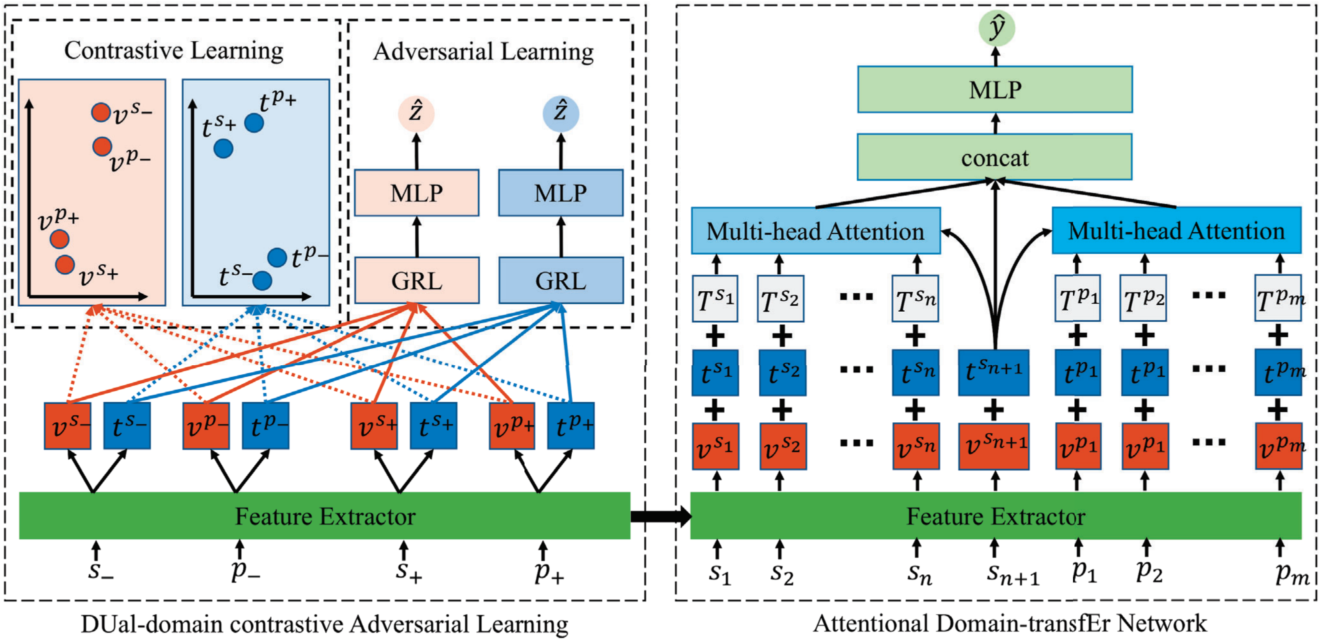

Contrastive Learning
Adversarial Learning
v^{s-}
v^{p-}
v^{p+}
v^{s+}
t^{p+}
t^{s+}
t^{s-}
t^{p-}
$\hat{z}$
MLP
GRL
Feature Extractor
s_-
p_-
s_+
p_+
$\hat{y}$
MLP
concat
Multi-head Attention
Multi-head Attention
T^{s_1} T^{s_2} ⋯ T^{s_n} T^{p_1} T^{p_2} ⋯ T^{p_m}
t^{s_1} t^{s_2} ⋯ t^{s_n} $t^{s_{n+1}}$ t^{p_1} t^{p_1} ⋯ t^{p_m}
v^{s_1} v^{s_2} ⋯ v^{s_n} $v^{s_{n+1}}$ v^{p_1} v^{p_1} ⋯ v^{p_m}
Feature Extractor
s_1 s_2 s_n s_{n+1} p_1 p_2 p_m
DUal-domain contrastive Adversarial Learning
Attentional Domain-transfEr Network

Fig. 5. Framework of our proposed method. s and p stand for storytelling and product respectively, v, t, and T are visual features, textual features, and timestamp respectively. In the first stage, we randomly select two pairs, which are denoted as $\{s_+, p_+\}$ and $\{s_-, p_-\}$ respectively, to pre-train the visual and textual feature extractors via contrastive adversarial learning. GRL stands for Gradient Reversal Layer and $\hat{z}$ is the domain label. In the second stage, we use a user's historical storytelling sequence $\{s_1, \ldots, s_n\}$ and historical product sequence $\{p_1, \ldots, p_m\}$ to predict his/her preference on the target storytelling item s_{n+1}, the predicted preference is denoted by $\hat{y}$.

and storytelling sequence, which means that user behaviors in different domains might reflect similar user preferences. From a global perspective, we calculate the distribution of categories in different domains in the SRD-2 dataset. We select the results of the top-20 popular categories in the storytelling domain, which are shown in Fig. 4. It can be observed that popular categories in the product domain may not be that popular in the storytelling domain, such as women's shoes, tableware, etc. This phenomenon reveals that user behavior patterns are different across the two domains.

We summarize our findings as follows.

- There is a large domain gap between product and storytelling.
- User behavior patterns are quite different in the storytelling domain and the product domain. However, there are some common interests between two domains.
- The sequential order information may be important since the sequence behaviors in the product domain reflect the short-term interest.

The above findings motivate us to design the following methods for cross-domain storytelling recommendation.

III. Our Method

In this section, we describe each part of our proposed ADAPT in detail. First, we introduce some notations. We denote the user clicked storytelling sequence and product sequence as $\mathbb{S} = \{s_1, s_2, ..., s_n\}$ and $\mathbb{P} = \{p_1, p_2, ..., p_m\}$ respectively, where n and m denote the length of either sequence. We use v to represent visual feature that is extracted from the cover picture, t to represent textual feature that is extracted from the title, and T to represent timestamp.

A. Feature Extractor

New storytellings are generated from different sources at the speed of hundreds of thousands per day, which leads to the *item cold-start* problem. To address it, we concentrate on how to leverage the feature of cover pictures and titles to make storytelling recommendations in this paper. Specifically, we have the visual feature vector $\hat{v}^{s_i} \in \mathbf{R}^{512}$ from its cover picture by a pre-trained CNN model for each storytelling s_i. For computational efficiency in the downstream task, we project $\hat{v}^{s_i}$ into a lower-dimensional space by learning an embedding matrix E_v, i.e. $v^{s_i} = E_v \hat{v}^{s_i}$, where $v^{s_i} \in \mathbf{R}^d$ is the visual feature representation of s_i and d is the embedding size.

In addition, we obtain the word embeddings $\{t^{s_i}_{w_1}, t^{s_i}_{w_2}, ..., t^{s_i}_{w_{K_i}}\}$ for each word in the title of storytelling s_i by look up from the word embedding matrix, where K_i is the number of words and the word embedding matrix is trained out. Then we perform self-attention to form textual feature representation t^{s_i} by imposing high weights on important words in the title. For the k-th word embedding $t^{s_i}_{w_k} \in \mathbf{R}^d$, the attention score is computed by a two-layer network,

$$\alpha(i,k) = W_\alpha \tanh(W_{1\alpha} t^{s_i}_{w_k} + b_{1\alpha}) \tag{2}$$

where $W_\alpha \in \mathbf{R}^{d \times d}, W_{1\alpha} \in \mathbf{R}^{d \times d}, b_{1\alpha} \in \mathbf{R}^d$ are trainable parameters. The attention score for each word is normalized as:

$$\tilde{\alpha}(i,k) = \frac{\exp(\alpha(i,k))}{\sum_{k=1}^{K_i} \exp(\alpha(i,k))} \tag{3}$$

Finally, the textual feature of storytelling s_i is

$$t^{s_i} = \sum_{k=1}^{K_i} \tilde{\alpha}(i,k) t^{s_i}_{w_k} \tag{4}$$

The procedure of extracting visual feature representation v^{p_i} of product p_i is the same with storytelling, where embedding matrix E_v is shared for mapping the visual feature representation of cross-domain into the same semantic space. For the textual feature of product p_i, we directly perform average pooling instead of attentional pooling since the title of the product is composed by keywords.

It is worth noting that unseen words may appear in the inference stage. Since such occurrences are rare in our datasets, we do not pay attention to this issue, and we use a default word embedding for the unseen.

B. DUal-domain contrastive Adversarial Learning

According to the analysis in Section II-C, we observe that there is a large domain gap between storytelling and product. To bridge the domain gap between storytelling and product, we propose a **DU**al-domain contrastive **A**dversarial **L**earning (named DUAL for short) method to pre-train the feature extractors for storytelling and product simultaneously, which is illustrated in Fig. 5. Our proposed DUAL model minimizes the domain gap between storytelling and product by relying on the characteristics that most storytelling items are usually associated with one or more products by the providers. For example, if the storytelling s_+ mounts two products p^1_+ and p^2_+, we can form two pairs $\{s_+, p^1_+\}$ and $\{s_+, p^2_+\}$ for this storytelling item. Given a pair of storytelling and product $\{s_+, p_+\}$, we also randomly select a pair $\{s_-, p_-\}$ from another storytelling item. And then we obtain the corresponding multi-modal features $\{v^{s_+}, t^{s_+}, v^{p_+}, t^{p_+}\}$ and $\{v^{s_-}, t^{s_-}, v^{p}\text{-}, t^{p_-}\}$ for two pairs respectively through the feature extractor. Inspired by the ranking-based recommender system [19], we impose dual-domain constraints on the feature representation in order to minimize the distance among the representations of storytelling and product of the same pairs. We use the function $d(x, y) = ||x - y||_2$ to compute the Euclidean distance where x and y are both d-dimensional vectors. Thus, the loss of dual-domain constraints on visual features is computed by

$$\begin{aligned} L_{d_v} =& f_h(d(v^{s_+}, v^{p_+}), d(v^{s_+}, v^{p_-})) \\ &+ f_h(d(v^{s_+}, v^{p_+}), d(v^{s_-}, v^{p_+})) \\ &+ f_h(d(v^{s_-}, v^{p_-}), d(v^{s_+}, v^{p_-})) \\ &+ f_h(d(v^{s_-}, v^{p_-}), d(v^{s_-}, v^{p_+})) \end{aligned} \tag{5}$$

where $f_h(x_1, x_2) = \max(0, \mu + x_1 - x_2)$, μ is the margin and is set to 1.0 in this paper. Likewise, we can compute the loss of dual-domain constraints L_{d_t} on textual features. Finally, we form the loss of dual-domain constraints on multi-modal features by $L_d = L_{d_t} + L_{d_v}$.

To further bridge the domain gap between the storytelling and product, adversarial learning [18], which is widely used in domain adaptation [20], [21] with great success, is applied for pre-training our feature extractors. Specifically, we design two domain classifiers D_v and D_t for distinguishing the visual features and the textual features in cross-domain, respectively. Our goal is to train feature extractors to produce domain-invariant representations so that the domain classifiers cannot reliably predict the domain of the encoded representation. The domain classifier D_v and D_t are implemented by Multi-Layer Perceptron (MLP) for predicting the label $\hat{z} \in \{0, 1\}$ of the visual feature and the textual feature, where 0 and 1 indicate the storytelling domain and the product domain respectively. The adversarial loss is defined by cross-entropy loss

$$L_{adv} = z \log \hat{z} + (1 - z) \log(1 - \hat{z}) \tag{6}$$

For simplicity, we denote the parameters used in feature extractors and domain classifiers as θ_E and θ_C, respectively. The procedure of learning multi-modal feature representation runs as a min-max game [18] of the two concurrent sub-processes:

$$\hat{\theta}_E = \arg\min_{\theta_E}(L_d(\theta_E) - L_{adv}(\hat{\theta}_C)) \tag{7}$$

$$\hat{\theta}_C = \arg\max_{\theta_C}(L_d(\hat{\theta}_E) - L_{adv}(\theta_C)) \tag{8}$$

During the implementation, we use Gradient Reversal Layer (GRL) [20], which has the same output as the identity function but reverses the gradient direction to perform a min-max game for adversarial learning simultaneously.

C. Attentional Domain-transfEr Network

Most existing cross-domain recommendation methods [1], [8], [9] are not suitable for our task since they are either not designed for multimedia content, or suffer from item *cold-start* problem. Because storytelling-domain user actions are more sparse than product-related user actions and user behavior patterns are largely diverse in cross-domain, we propose an **A**ttentional **D**omain-transf**E**r **N**etwork (ADEN) shown in Fig. 5, which can automatically select related items in the two domains to capture the user preferences dynamically. As user behavior patterns are quite different in the storytelling domain and the product domain, we leverage two multi-head attention models [16] to encode sequences $\mathbb{S}$ and $\mathbb{P}$ respectively. Since the multi-head attention is not aware of the order of the sequence, and the sequential behaviors in the product domain reflect the short-term interests, we use timestamp embedding to preserve the sequential order information so as to capture dynamic user preferences over time across the storytelling domain and the product domain. We first split the user's historical behaviors into multiple time blocks, where the length of each time block is one hour. Then we learn a timestamp embedding for each time block, and storytelling or product in the same time block will share the same timestamp embedding. Consequently, the multi-modal feature $f^{s_i} \in \mathbf{R}^d$ ($f^{p_i} \in \mathbf{R}^d$) is formed by summing up the visual feature v^{s_i} (v^{p_i}), the textual feature t^{s_i} (t^{p_i}), and the timestamp feature T^{s_i} (T^{p_i}) where v^{s_i} (v^{p_i}) and t^{s_i} (t^{p_i}) are extracted from our pre-trained feature extractors. Finally, we stack f^{s_i} (f^{p_i}) together into matrix $F^s \in \mathbf{R}^{n \times d}$ ($F^p \in \mathbf{R}^{m \times d}$) since we compute attention weight over each item simultaneously. The multi-head attention models evaluate the similarity between the target storytelling and sequence in the two domains respectively, to generate the domain-specific user representation in each domain. Taking multi-head attention used in the product domain as an example, multi-head attention first linearly projects $f^{s_{n+1}}$ and F^p into h subspaces with different trainable parameters respectively, and then applies attention h times in parallel. The independent attention outputs are concatenated to form the final user representation u_p in the product domain.

$$\begin{aligned} u_p &= [head_1; \ldots; head_h] \\ head_i &= \text{Attention}(f^{s_{n+1}}W_i^{Q_p}, F^pW_i^{K_p}, F^pW_i^{V_p}) \end{aligned} \tag{9}$$

where $W_i^{Q_p} \in \mathbf{R}^{d \times d_h}$, $W_i^{K_p} \in \mathbf{R}^{d \times d_h}$, and $W_i^{V_p} \in \mathbf{R}^{d \times d_h}$ are parameter matrices to be learned, $d_h = d/h$, h is the number of attention heads and set to 4. Similarly, we obtain the user representation u_s in the storytelling domain with another multi-head attention model. Here we adopt scaled dot-product attention mechanism,

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_h}})V \tag{10}$$

the Attention function use target storytelling as a query to compute similarity with each item in cross-domain in order to automatically select relevant items in two domains.

Note that there have been many works exploring the strategies of combining multiple representations, such as addition, element-wise product, or more complex function. In our datasets, we find that concatenation is the most efficient way. The concatenation of u_s, u_p and $f_{s_{n+1}}$ are fed into a multi-layer perceptron to predict whether user click or not. We use the binary cross-entropy loss as the objective to minimize:

$$L_{ctr} = y \log \hat{y} + (1 - y) \log(1 - \hat{y}) \tag{11}$$

where $y \in \{0, 1\}$ is the ground-truth that indicates whether the user clicks the storytelling or not in the training data.

IV. Experiments

In this section, we present our experiments in detail, including compared methods, implementation details, evaluation metrics, experimental results, and the corresponding analyses.

A. Compared Methods

For our contrastive experiments, we consider baselines from two aspects: single-domain and cross-domain, where single-domain methods only use the information in the storytelling domain. Representation of storytelling used in all baseline models is introduced in Section III-A.

1) Single-domain recommendation: As single-domain recommendation approaches, we consider the following:

- **MF** [22]: In our setting, we use a low-dimensional vector to represent user preference.
- **NCF** [23]: Neural collaborative filtering extends MF into a deep framework, which learns both user embedding and item representation with a shallow network (element-wise product of user and item) and a deep network (concatenation of user and item followed by several feed-forward layers).
- **YouTubeNet** [24]: YouTubeNet is a deep model for video recommendation on YouTube, which gets user representations by simply averaging the item representations in the user behavior sequence.
- **THACIL** [25]: It is a self-attention based method for the micro-video recommendation, which utilizes a forward multi-head self-attention layer to capture the long-term correlation within user behaviors. It also uses two-level (i.e. item and category) attention layer to model the fine-grained profiling of the user interest.
- **DIN** [26]: Deep interest network uses the attention model to adaptively learn the representation of user preferences from historical behaviors with respect to a certain item instead of averaging pooling.
- **DIEN** [27]: Deep interest evolution network is an improved version of DIN, which proposes an interest evolving layer on top of DIN model to capture user's dynamic interest over time.
- **BST** [28]: A recent state-of-the-art method on modeling user interest based on sequential behavior, which leverages the Transformer [16] layer with time information for recommendation.

2) Cross-domain recommendation: As cross-domain recommendation approaches, we consider the following:

- **CMF** [1]: Collective matrix factorization factorizes matrix in different domains jointly, which transfers knowledge by sharing user latent factors.
- **CoNet** [8]: Collaborative cross network extends CMF into a deep framework for cross-domain recommendation, which employs a cross-stitch network [29] to enable knowledge transfer.
- **Cross-BST**: We extend BST [28] into cross-domain, where we directly treat storytelling item as a normal product in order to form one behavior sequence instead of using two cross-domain behavior sequences like in our proposed ADEN.
- **π-Net** [30]: A recent state-of-the-art method for cross-domain sequential recommendation, which uses RNNs to encode sequences and a gating mechanism to transfer information between sequences in cross-domain.

B. Implementation Details

All the methods, including ours and the compared baselines, are implemented with TensorFlow and running on an NVIDIA GTX 1080Ti graphical processing unit. The embedding size d is 64. Here an implementation challenge is that different users have different numbers of historically interacted storytellings and products, while TensorFlow requires all training instances of a batch must be of the same length. To overcome this challenge, we use mask trick, which uses zero padding to ensure all instances of a batch have the same length. In all sequence-based methods, the historical length is set to 20. The batch size is set to 128. AdamW optimizer [31] is adopted for training. Learning rate is 0.0001. To prevent over-fitting, L_2 regularization coefficient of all models is tuned in the range of $[10^{-7}; 10^{-6}; ...; 1]$ for best performance. In our model, L_2 regularization coefficient is 0.01. Dropout rate is set to 0.1.

TABLE IV
PERFORMANCE OF DIFFERENT METHODS ON THE TWO DATASETS. **BOLD** INDICATES THE BEST AND UNDERLINE INDICATES THE SECOND BEST. * INDICATES THAT THE IMPROVEMENT OF THE BEST RESULT IS STATISTICALLY SIGNIFICANT COMPARED WITH THE SECOND BEST RESULT WITH $p < 0.01$.

Group	Method	SRD-1			SRD-2		
		AUC	HR@5	NDCG@5	AUC	HR@5	NDCG@5
Single-domain	MF	0.5738	0.4486	0.2778	0.5756	0.4395	0.2719
	NCF	0.5856	0.4646	0.2877	0.5926	0.4605	0.2853
	YouTubeNet	0.6045	0.4895	0.3049	0.6216	0.5022	0.3131
	THACIL	0.6105	0.5018	0.3130	0.6256	0.5063	0.3165
	DIN	0.6131	0.5023	0.3133	0.6296	0.5141	0.3213
	DIEN	0.6152	0.5049	0.3153	0.6319	0.5172	0.3235
	BST	0.6180	0.5095	0.3189	0.6370	0.5245	0.3281
Cross-domain	CMF	0.5762	0.4535	0.2797	0.5801	0.4436	0.2747
	CoNet	0.5827	0.4616	0.2866	0.5895	0.4570	0.2831
	Cross-BST	0.6329	0.5309	0.3327	0.6605	0.5580	0.3519
	π-Net	0.6426	0.5478	0.3432	0.6665	0.5696	0.3597
	ADAPT	**0.6499***	**0.5573***	**0.3506***	**0.6742***	**0.5815***	**0.3684***

TABLE V
NUMBER OF FLOATING POINT OPERATIONS (FLOPS) OF ADAPT AND π-NET. 'ALL' INDICATES THE COMPLEXITY OF THE ENTIRE MODEL, AND 'SEQ' INDICATES THE COMPLEXITY OF THE SEQUENCE ENCODER ONLY. THE FLOPS OF PRE-TRAINING THE DUAL IN THE ADAPT ARE ALREADY INCLUDED FOR FAIR COMPARISON.

Method	Training cost ($\times 10^6$)		Inference cost ($\times 10^6$)	
	All	Seq	All	Seq
ADAPT	69.07	2.08	38.97	0.73
π-Net	80.53	14.44	41.78	4.22

The MLP used in DUAL and ADEN is set to $64 \rightarrow 1$ and $256 \rightarrow 128 \rightarrow 1$ respectively. All methods are trained out for best performance.

C. Evaluation Metrics

To evaluate the overall performance of click-through rate prediction by different methods, we adopt the widely used Area Under Curve (AUC) as metric, which is defined as:

$$AUC = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{V}_u^+||\mathcal{V}_u^-|} \sum_{s_i \in \mathcal{V}_u^+} \sum_{s_j \in \mathcal{V}_u^-} \delta(\hat{y}_{u,s_i} > \hat{y}_{u,s_j}) \tag{12}$$

where $\hat{y}_{u,s_i}$ is the predicted score that a user u may click a storytelling s_i in the test set, $|\mathcal{U}|$ is the number of all users, $\mathcal{V}_u^+$ and $\mathcal{V}_u^-$ consist of the items that the user u actually clicked and actually not clicked, respectively, $\delta(\cdot)$ is the indicator function. In addition, the click-through rate prediction results can be used for generating top-k recommendations. So we further choose Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) [32] to evaluate the recommended sets quantitatively, where k is set to 5.

D. Performance Comparison

Table IV presents the experimental results for measuring performance of different methods. From the results, we have the following observations.

Firstly, our method ADAPT consistently and significantly outperforms all the baselines in terms of all metrics on both datasets. For example, ADAPT outperforms the best baseline by about 0.95% on SRD-1 dataset and 1.19% on SRD-2 dataset on the HR@5, respectively. We run 5 times for each method with random initializations, and perform Student's t-test between our method and the best baseline. The results indicate that the improvements on all metrics are statistically significant ($p < 0.01$) compared to the best baseline. In addition, our proposed ADAPT is computationally more efficient than π-Net in the training and inference stages. We show the number of floating-point operations (FLOPs) of our model as well as π-Net in Table V. It can be observed that ADAPT has lower complexity than π-Net especially in the part of sequence encoder. As for the timing results, our ADAPT needs 1.8 days to train and π-Net needs 5.8 days on the SRD-2 dataset. ADAPT is at least $3\times$ faster than π-Net. It is worth noting that the timing is affected by multiple factors including CPU/GPU communication and disk I/O.

Secondly, sequence-based methods (e.g., YouTubeNet, THACIL, DIN, DIEN, BST, Cross-BST, π-net, ADAPT) generally perform better than collaborative filtering-based methods (e.g., MF, NCF, CMF, CoNet). This is not surprising since sequence-based methods leverage the user behavior sequence to represent user preferences instead of a low-dimensional vector used in CF-based methods.

Thirdly, the performance of cross-domain recommendation is better than the single-domain recommendation in terms of the same kinds of methods. For example, BST and Cross-BST are both sequence-based methods. Cross-BST outperforms BST because Cross-BST transfers knowledge from the product domain into the storytelling domain by introducing the user behavior sequences in the product domain.

TABLE VI
RESULTS OF ABLATION STUDY.

Group	Method	SRD-1			SRD-2		
		AUC	HR@5	NDCG@5	AUC	HR@5	NDCG@5
Storytelling-domain	V	0.6139	0.5034	0.3144	0.6328	0.5180	0.3242
	T	0.6131	0.5025	0.3133	0.6330	0.5191	0.3247
	V+T	0.6197	0.5120	0.3198	0.6373	0.5244	0.3288
Product-domain	V	0.6003	0.4864	0.3026	0.6241	0.5036	0.3149
	T	0.5984	0.4818	0.2992	0.6227	0.5040	0.3148
	V+T	0.6069	0.4932	0.3072	0.6306	0.5128	0.3206
Cross-domain	V w/o DUAL	0.6295	0.5261	0.3290	0.6555	0.5510	0.3469
	V w. DUAL	0.6391	0.5413	0.3394	0.6608	0.5599	0.3535
	T w/o DUAL	0.6265	0.5247	0.3282	0.6523	0.5465	0.3442
	T w. DUAL	0.6429	0.5473	0.3434	0.6677	0.5708	0.3607
	ADAPT w/o DUAL	0.6360	0.5365	0.3359	0.6642	0.5678	0.3587
	ADAPT	**0.6499**	**0.5573**	**0.3506**	**0.6742**	**0.5815**	**0.3684**

E. Ablation Study

We conduct a comprehensive ablation study to investigate the impact of each component in our proposed method, which is shown in Table VI. We denote the models as V, T, and V+T which mean the results of visual feature, textual feature and the combination of visual and textual feature, respectively. Note that timestamp is used always. We also evaluate the effectiveness of DUAL by using or not using it.

1) Effect of individual modality: Firstly, we observe that the performance of V and T in the storytelling domain is close, and the performance of V is slightly better than T in the product domain on both datasets. Secondly, we find that the NDCG [32] performance of V+T (i.e., multi-modal feature) outperforms the individual feature V or T in each domain on both datasets. In summary, the visual and textual features are almost equally crucial for storytelling recommendation, and the multi-modal feature can achieve the best performance.

2) Effect of ADEN: It can be observed that the performance of the product domain is worse than the storytelling domain in terms of all features on both datasets. This phenomenon empirically demonstrates the difference of user behavior patterns across domains and a large domain gap between the storytelling domain and the product domain. Moreover, the performance of cross-domain outperforms each single-domain by a large margin in terms of all modality features since our proposed ADEN can select related items in each domain to identify user preference. In conclusion, storytelling domain contributes larger than product domain for storytelling recommendation, and leveraging the cross-domain information can improve the recommendation performance by aggregating user preference in different domains dynamically.

3) Effect of DUAL: The performance of models with DUAL consistently outperforms models without DUAL in terms of all modality features on both datasets. This confirms that DUAL can improve the recommendation performance by reducing the domain gap between storytelling and product. Moreover, we find that the improvement of V+DUAL over V is lower than T+DUAL over T. This might be the reason that the domain gap in the title is more significant than in the image. Thus, our proposed DUAL is helpful to improve recommendation performance.

TABLE VII
PERFORMANCE OF RECOMMENDING STORYTELLING TO PRODUCTS.

Method	SRD-1			SRD-2		
	AUC	HR@5	NDCG@5	AUC	HR@5	NDCG@5
w/o DUAL	0.9055	0.2020	0.2698	0.9127	0.2637	0.2939
w. DUAL	0.9196	0.3294	0.3221	0.9249	0.3730	0.3405

TABLE VIII
AVERAGE SUBJECTIVE RATINGS OF RECOMMENDING STORYTELLING TO PRODUCTS.

Method	Matching	Interesting	Purchasing
w/o DUAL	3.12	2.78	2.47
w. DUAL	3.82	3.38	3.04

F. Results of Online A/B Test

We have deployed the proposed method in the Taobao service. In the period of Nov. 23–26, 2020, we perform an online A/B test that compares ADAPT with CrossBST (we do not deploy π-Net in the online A/B test due to its high complexity). Quantitatively, ADAPT increases the number of videos viewed by users by 9 million on the basis of 220 million (i.e., 4.3% increase); also, ADAPT increases the amount of dwell time by 4.9%, which enhances user stickiness to the service.

G. Recommending Storytelling to Product

We have considered the recommendation of storytelling items for normal users. In e-commerce platforms, there is another requirement to recommend storytelling for product providers or merchants. This is because currently most of

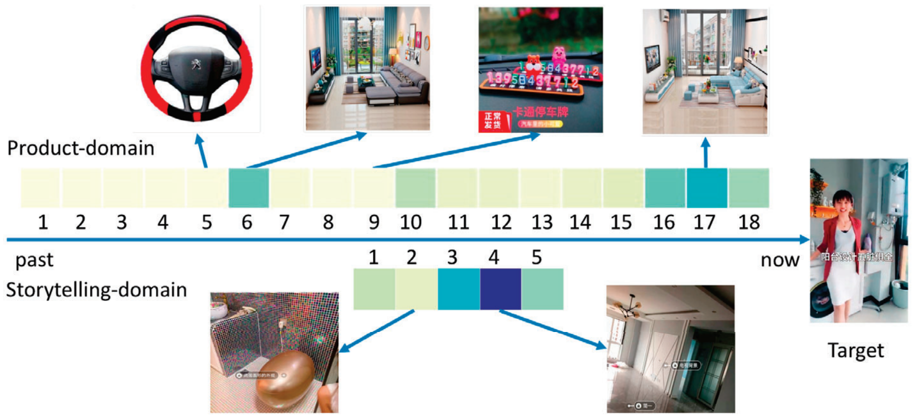


Fig. 6. Example of the learned attention weights in ADEN. Each number represents an item in the corresponding domain, and the color of each item is to denote its weight (the darker the color, the higher the weight).

the products do not have related storytelling, and mounting a storytelling on a product can inspire users to purchase the product. We consider this requirement in this subsection.

The task of recommending storytelling to product can be seen as a matching problem. Our proposed ADAPT can be used for this task, because we have extracted features of both storytelling and product, and the features are put into the same latent space. Our proposed DUAL is especially useful because we make the features better aligned through the adversarial learning.

Specifically, we use the trained feature extractors to compute the feature vector of every storytelling and product. For each storytelling, in the feature space, we find out its nearest neighbors of products, where the distance is simply Euclidean.

For objective evaluation, we use the storytellings that mount products as search results (“ground truth”) and the mounted products as queries. For comparison, the widely used metrics, including AUC, HR, and NDCG, are calculated to evaluate different methods. Note that we take the top-500 recalled storytellings to calculate the AUC metric for an overall evaluation. HR and NDCG are used for evaluating the top-k results where k is set to 5. The results are shown in Table VII. Firstly, it can be observed that DUAL helps improve the results on both datasets. Secondly, comparing the two datasets, the results of SRD-2 are better than those of SRD-1. SRD-2 has more storytellings and products so it may appear more difficult to recommend storytelling for each product. Nonetheless, SRD-2 has much more instances to train the ADEN, as well as more pairs to pre-train the feature extractors. It reveals that the more data can help train the network better.

We also conduct a subjective evaluation to compare the method with and without DUAL. We randomly select 100 products and recommend the nearest storytelling to each product. Then we invite 24 volunteers to assess the quality of recommendations. Each volunteer is asked to choose 5 most favorite products to assess. Each product and the two recommendations (with and without DUAL) are shown to the subject at the same time. The assessment is performed at three aspects: matching degree (how matching is the storytelling to the product), interesting degree (how interesting in the storytelling itself), and purchasing probability (how probable to purchase the product). The rating is given in 5-score metric and the higher the better. We then collect the ratings and calculate their average, as shown in Table VIII. It is confirmed, by the t-test with $p < 0.05$, that using DUAL achieves significantly better results than not using DUAL. These results again demonstrate the advantage of DUAL.

H. Case Study

In this subsection, we conduct experiments to further understand how our proposed method works. To investigate how the attention model in ADAPT works, we randomly select one interaction and visualize the attention weight of user behavior sequence in storytelling domain and product domain respectively. As shown in Fig. 6, the topic of target storytelling describes how to decorate house. It can be observed that the attention model aligns the high weight and low weight to storytelling #4 and #2 respectively since storytelling #4 also describes how to decorate house but storytelling #2 describes closest tool. It is worth noting that the attention model leads to high weight and low weight on products {#6, #17} and products {#5, #9} respectively. Products {#5, #9} are about car and products {#6, #17} are about sofa. It is not surprising to give low weight on products {#5, #9} since the car is not very relevant to the decoration of house. According to our investigation of the case, we find that the product mounted on the target storytelling is about sofa. Thus, the reason might be that our proposed DUAL bridges the domain gap between target storytelling and sofa so that the attention model leads to high weight on products {#6, #17}.

In addition, we observe some cases to analyze under which scenarios our proposed ADAPT may fail. To this end, we select two instances where the predicted ranking score is far from the true label, which are shown in Table IX. We denote the models as ADAPT$_s$, ADAPT$_p$, and ADAPT, to represent the results of using user behavior sequences in the storytelling domain, in the product domain, and (by default) in both domains, respectively. For case #1, we observe that the predicted ranking scores of ADAPT$_s$ are close to the true label, but ADAPT$_p$ and ADAPT fail. In this case, the target storytelling describes notes for beginners in learning to

TABLE IX
EXAMPLES OF FAILURE CASES.

Case	Label	Method	Ranking Score	
			w/o DUAL	w. DUAL
#1	1	ADAPT$_s$	0.6273	0.6352
		ADAPT$_p$	0.2109	0.2056
		ADAPT	0.2844	0.2979
#2	0	ADAPT$_s$	0.4218	0.6023
		ADAPT$_p$	0.2452	0.7820
		ADAPT	0.3357	0.7431

play the instrument, and the user has recently clicked some storytellings about playing the other instruments. However, the user's historically clicked product items are all girl's skirts and pajamas. Since the target storytelling is not related to the user's items in the product domain, ADAPT$_p$ predicts low ranking scores for the target storytelling, which then affects the predicted ranking scores of ADAPT. For this case, a feasible solution is to dynamically adjust the weights of user interest representations from different domains. For case #2, we observe that the predicted ranking score of all the methods with DUAL are worse than their counterparts without DUAL. The title of the target storytelling is: "how would you react when your mother has a second child?" The mounted products in the target storytelling are lipstick and makeup tools. When people first see such a title, they usually think it is for maternal products. Since the user's items in the product domain and storytelling domain are both not related to maternal products, all methods without DUAL tend to predict low ranking scores. However, the proposed DUAL explicitly aligns the feature representations of the storytelling and its mounted products, and leads to high ranking scores. Fortunately, such highly imaginary titles are rare, and most of the storytelling titles match the content intuitively.

V. RELATED WORK

Cross-Domain Recommendation: Cross-domain recommendation algorithms [1]–[9], [30], [33], [34] have proven to be beneficial for alleviating *cold-start* and data sparsity problems by leveraging auxiliary data from source domain to improve recommendation accuracy in the target domain. Approaches of cross-domain recommendation can be broadly categorized into two types: traditional methods and deep learning-based methods. Traditional methods refer to transferring knowledge among different domains using shallow models. For example, Ajit *et al.* proposed Collective Matrix Factorization (CMF), which extends Matrix Factorization by sharing the user latent factors, to transfer knowledge within different domains. Loni *et al.* [7] modeled user preference across difference domains by factorization machines (FM) [35] to generate recommendation in the target domain. Compared with traditional methods, deep learning is better to learn high-level and complex information within multiple domains. For example, Elkahky *et al.* [33] introduced a multi-view deep learning model to make recommendation by combining rich user features from multiple domains. Hu *et al.* proposed Collaborative cross Networks (CoNet) to enable dual knowledge transfer across domains based on a cross-stitch network [29]. More recently, Kanagawa *et al.* [36] used Domain Separation Network [37] to learn domain-specific and domain-invariant user representation. Wang *et al.* [38] extended neural collaborative filtering into cross-domain recommendations for alleviating user cold-start problem in e-commerce by sharing user representation. Ma *et al.* [30] proposed a parallel information-sharing network named π-Net, which enables information transfer in cross-domain sequence at each timestamp, for shared-account cross-domain sequential recommendation. Although most existing methods achieve acceptable performance in many applications, they are not suitable for cold-start problem and ignore the sequential characteristics.

Multimedia Recommendation: The significance of multimedia recommendation has led to the great attention of researchers [13]–[15], [25], [39]–[46]. For example, Gao *et al.* [43] proposed a dynamic recurrent neural network to model users' dynamic interests for video recommendation. Zhang *et al.* [41] leveraged the tag of images and deep feature learned by CNN to represent user interest for image recommendation. Han *et al.* [42] took advantage of the intrinsic motion information (dance style) to recommend dance video. Chen *et al.* [25] proposed a temporal hierarchical attention at category- and item-level network for micro-video recommendation. In this paper, we study recommending storytelling, a new type of content in e-commerce, and address this requirement by a cross-domain personalized recommendation method.

E-Commerce Recommendation: The significance of e-commerce recommendation has led to the great attention from both industry and academia [26], [28], [47]–[52]. For example, Zhou *et al.* [26] proposed an attention model named DIN, which is to adaptively learn the representation of user interests from historical behaviors in e-commerce platform. Furthermore, they proposed DIEN [48] and MIMN [52] to capture dynamic user interest and long-term user interest for e-commerce business respectively. Huang *et al.* [51] proposed a graph multi-scale pyramid network to exploit users' latent behavior patterns for online purchase prediction. Lin *et al.* [49] proposed a cross-platform recommendation model by adopting user behaviors on product in social platform. Existing e-commerce recommendation approaches are usually proposed for recommending product. The recommendation of storytelling is more challenging than the recommendation of product because storytelling and product have different data formats as we have analyzed before. In addition, since storytelling is a new type of content, its related user actions are much less, more sparse than product-related user actions. To address the challenges, we propose ADAPT as a cross-domain recommendation solution for storytelling, and we propose DUAL to bridge the domain gap between product and storytelling.

VI. CONCLUSION

In this paper, we investigate the task of storytelling recommendation in e-commerce. We address the task by a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3054525, IEEE Transactions on Multimedia

cross-domain approach. Specifically, we propose an attentional domain-transfer network to identify user preferences in the storytelling domain and the product domain. To bridge the gap between the two domains, we propose a dual-domain contrastive adversarial learning method to pre-train the feature extractors. Our experiments on two industrial datasets demonstrate the effectiveness of our proposed method compared with state-of-the-art methods.

Storytelling recommendation is an emerging task in e-commerce and remains largely unexplored. We may utilize much richer information of storytelling items, such as audio and motion features. Another interesting and challenging task is to recommend storytelling items to product providers according to their branding requirements. Moreover, it is possible and very interesting to apply ADAPT as a cross-domain method for other recommendation tasks, like product recommendation, which will be studied in future work.

REFERENCES

[1] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *KDD*, 2008, pp. 650–658.

[2] B. Li, Q. Yang, and X. Xue, "Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction," in *IJCAI*, 2009, pp. 2052–2057.

[3] F. Zhuang, P. Luo, H. Xiong, Y. Xiong, Q. He, and Z. Shi, "Cross-domain learning from multiple sources: A consensus regularization perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 12, pp. 1664–1678, 2009.

[4] B. Li, X. Zhu, R. Li, C. Zhang, X. Xue, and X. Wu, "Cross-domain collaborative filtering over time," in *IJCAI*, 2011, pp. 2293–2298.

[5] S. Gao, H. Luo, D. Chen, S. Li, P. Gallinari, and J. Guo, "Cross-domain recommendation via cluster-level latent factor model," in *ECML PKDD*, 2013, pp. 161–176.

[6] L. Hu, J. Cao, G. Xu, L. Cao, Z. Gu, and C. Zhu, "Personalized recommendation via cross-domain triadic factorization," in *WWW*, 2013, pp. 595–606.

[7] B. Loni, Y. Shi, M. Larson, and A. Hanjalic, "Cross-domain collaborative filtering with factorization machines," in *ECIR*, 2014, pp. 656–661.

[8] G. Hu, Y. Zhang, and Q. Yang, "Conet: Collaborative cross networks for cross-domain recommendation," in *CIKM*, 2018, pp. 667–676.

[9] C. Gao, X. Chen, F. Feng, K. Zhao, X. He, Y. Li, and D. Jin, "Cross-domain recommendation without sharing user-relevant data," in *WWW*, 2019, pp. 491–502.

[10] T. Mei, B. Yang, X.-S. Hua, and S. Li, "Contextual video recommendation by multimodal relevance and user feedback," *ACM Transactions on Information Systems*, vol. 29, no. 2, pp. 1–24, 2011.

[11] Q. Zhu, M.-L. Shyu, and H. Wang, "Videotopic: Content-based video recommendation using a topic model," in *IEEE International Symposium on Multimedia*, 2013, pp. 219–222.

[12] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana, "Content-based video recommendation system based on stylistic visual features," *Journal on Data Semantics*, vol. 5, no. 2, pp. 99–113, 2016.

[13] C. Lei, D. Liu, W. Li, Z.-J. Zha, and H. Li, "Comparative deep learning of hybrid representations for image recommendations," in *CVPR*, 2016, pp. 2545–2553.

[14] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention," in *SIGIR*, 2017, pp. 335–344.

[15] X. Chen, D. Liu, C. Lei, R. Li, Z.-J. Zha, and Z. Xiong, "BERT4SessRec: Content-based video relevance prediction with bidirectional encoder representations from transformer," in *MM*, 2019, pp. 2597–2601.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[19] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *UAI*, 2009, pp. 452–461.

[20] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 513–520.

[21] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[22] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[23] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *WWW*, 2017, pp. 173–182.

[24] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *ResSys*, 2016, pp. 191–198.

[25] X. Chen, D. Liu, Z.-J. Zha, W. Zhou, Z. Xiong, and Y. Li, "Temporal hierarchical attention at category-and item-level for micro-video click-through prediction," in *MM*, 2018, pp. 1146–1153.

[26] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep interest network for click-through rate prediction," in *KDD*, 2018, pp. 1059–1068.

[27] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, and K. Gai, "Deep interest evolution network for click-through rate prediction," in *AAAI*, 2019, pp. 5941–5948.

[28] Q. Chen, H. Zhao, W. Li, P. Huang, and W. Ou, "Behavior sequence transformer for e-commerce recommendation in alibaba," in *DLP-KDD*, 2019, pp. 1–4.

[29] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *CVPR*, 2016, pp. 3994–4003.

[30] M. Ma, P. Ren, Y. Lin, Z. Chen, J. Ma, and M. d. Rijke, "π-net: A parallel information-sharing network for shared-account cross-domain sequential recommendations," in *SIGIR*, 2019, pp. 685–694.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[32] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," in *SIGIR*, 2017, pp. 243–250.

[33] A. M. Elkahky, Y. Song, and X. He, "A multi-view deep learning approach for cross domain user modeling in recommendation systems," in *WWW*, 2015, pp. 278–288.

[34] J. Lian, F. Zhang, X. Xie, and G. Sun, "CCCFNet: A content-boosted collaborative filtering neural network for cross domain recommender systems," in *WWW*, 2017, pp. 817–818.

[35] S. Rendle, "Factorization machines," in *ICDM*, 2010, pp. 995–1000.

[36] H. Kanagawa, H. Kobayashi, N. Shimizu, Y. Tagami, and T. Suzuki, "Cross-domain recommendation via deep domain adaptation," in *ECIR*, 2019, pp. 20–29.

[37] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *NIPS*, 2016, pp. 343–351.

[38] H. Wang, D. Amagata, T. Maekawa, T. Hara, H. Niu, K. Yonekawa, and M. Kurokawa, "Preliminary investigation of alleviating user cold-start problem in e-commerce with deep cross-domain recommender system," in *WWW*, 2019, pp. 398—-403.

[39] Z. Zhao, Q. Yang, H. Lu, T. Weninger, D. Cai, X. He, and Y. Zhuang, "Social-aware movie recommendation via multimodal network learning," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 430–440, 2017.

[40] L. Sun, X. Wang, Z. Wang, H. Zhao, and W. Zhu, "Social-aware video recommendation for online social groups," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 609–618, 2016.

[41] J. Zhang, Y. Yang, L. Zhuo, Q. Tian, and X. Liang, "Personalized recommendation of social images by constructing a user interest tree with deep features and tag trees," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2762–2775, 2019.

[42] T. Han, H. Yao, C. Xu, X. Sun, Y. Zhang, and J. J. Corso, "Dancelets mining for video recommendation based on dance styles," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 712–724, 2016.

[43] J. Gao, T. Zhang, and C. Xu, "A unified personalized video recommendation via dynamic recurrent neural networks," in *MM*, 2017, pp. 127–135.

[44] X. Du, H. Yin, L. Chen, Y. Wang, Y. Yang, and X. Zhou, "Personalized video recommendation using rich contents from videos," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 3, pp. 492–505, 2020.

[45] Y. Li, M. Liu, J. Yin, C. Cui, X.-S. Xu, and L. Nie, "Routing micro-videos via a temporal graph-guided recommendation system," in *MM*, 2019, pp. 1464–1472.

[46] X. Chen, D. Liu, Z. Xiong, and Z.-J. Zha, "Learning and fusing multiple user interest representations for micro-video and movie recommendations," *IEEE Transactions on Multimedia*, vol. 23, pp. 484–496, 2021.

[47] J. Wang, P. Huang, H. Zhao, Z. Zhang, B. Zhao, and D. L. Lee, "Billion-scale commodity embedding for e-commerce recommendation in alibaba," in *KDD*, 2018, pp. 839–848.

[48] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, and K. Gai, "Deep interest evolution network for click-through rate prediction," in *AAAI*, 2019, pp. 5941–5948.

[49] T.-H. Lin, C. Gao, and Y. Li, "Cross: Cross-platform recommendation for social e-commerce," in *SIGIR*, 2019, pp. 515–524.

[50] Y. Gu, Z. Ding, S. Wang, and D. Yin, "Hierarchical user profiling for e-commerce recommender systems," in *WSDM*, 2020, pp. 223–231.

[51] C. Huang, X. Wu, X. Zhang, C. Zhang, J. Zhao, D. Yin, and N. V. Chawla, "Online purchase prediction via multi-scale modeling of behavior dynamics," in *KDD*, 2019, pp. 2613–2622.

[52] Q. Pi, W. Bian, G. Zhou, X. Zhu, and K. Gai, "Practice on long sequential user behavior modeling for click-through rate prediction," in *KDD*, 2019, pp. 2671–2679.

Xusong Chen received the B.S. degree in computer science from Chang An University, Xi'an, China, in 2016. He is currently working toward the Ph.D. degree with the University of Science and Technology of China, Hefei, China. His research interests include recommender system, multimedia data mining, and information retrieval. He and his collaborators were winners of two technical challenges held in ACM MM 2019 and ACM MM 2018.

Chenyi Lei received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2013 and 2016, respectively. He is now an algorithm expert at Alibaba Group, Hangzhou, China, and meanwhile a Ph.D. student at USTC since 2019, advised by Prof. Houqiang Li. His current research interests include multimedia content analysis, recommender system, information retrieval, and machine learning. He has more than 10 publications in top conferences such as CVPR, WWW, and AAAI, and journals including the IEEE Transactions on Multimedia, and Complexity. Moreover, he has served as the PC Member for several top conferences including AAAI, IJCAI, and CVPR.

Dong Liu (M'13–SM'19) received the B.S. and Ph.D. degrees in electrical engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. He was a Member of Research Staff with Nokia Research Center, Beijing, China, from 2009 to 2012. He joined USTC in 2012 and became a Professor in 2020.

His research interests include image and video processing, coding, analysis, and data mining. He has authored or co-authored more than 100 papers in international journals and conferences. He has 19 granted patents. He has several technical proposals adopted by international and domestic standardization groups. He received the 2009 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Paper Award and the VCIP 2016 Best 10% Paper Award. He and his students were winners of several technical challenges held in ICCV 2019, ACM MM 2019, ACM MM 2018, ECCV 2018, CVPR 2018, and ICME 2016. He is a Senior Member of CCF and CSIG, an elected member of MSA-TC of IEEE CAS Society. He serves or had served as the Chair of IEEE Future Video Coding Study Group, a Publicity Co-Chair for ICME 2021, and a Registration Co-Chair for ICME 2019.

Guoxin Wang received the B.S. and M.S. degrees from Zhejiang University and is working toward the Ph.D. degree at Zhejiang University. He is currently an algorithm researcher at Alibaba Group. His current research interests include recommender systems, multimedia retrieval, and multimedia content analysis.

Haihong Tang is now a senior staff engineer & director in the Search and Recommendation Business Unit of Alibaba Group, leading the Multimedia Content Algorithm Team in multimedia content analyzing, search and recommendation. Her research interests include information retrieval, data mining, recommendation system, natural language processing, big multimedia data search, understanding, and mining, as well as pattern recognition and machine learning. She has published papers in SIGKDD, WWW, SIGIR, AAAI, IJCAI, ICDE, WSDM.

Zheng-Jun Zha (M'08) received the B.E. and Ph.D. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. He was a Researcher with the Hefei Institutes of Physical Science, Chinese Academy of Sciences, from 2013 to 2015, a Senior Research Fellow with the School of Computing, National University of Singapore (NUS), from 2011 to 2013, and a Research Fellow with NUS from 2009 to 2010. He is currently a Full Professor with the School of Information Science and Technology, USTC, and the Executive Director of the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application.

His research interests include multimedia analysis, retrieval and applications, and computer vision. He has authored or coauthored more than 100 papers in these areas with a series of publications on top journals and conferences. He was a recipient of multiple paper awards from prestigious multimedia conferences, including the Best Paper Award and Best Student Paper Award in ACM Multimedia. He serves as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Houqiang Li (SM'12–F'21) received the B.S., M.Eng., and Ph.D. degrees in electronic engineering from the University of Science and Technology of China (USTC), Hefei, China, in 1992, 1997, and 2000, respectively.

He is currently a Professor with the Department of Electronic Engineering and Information Science of USTC. His research interests include video coding and communication, multimedia search, image/video analysis. He has authored and co-authored over 100 papers in journals and conferences. He served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2010 to 2013, and has been with the Editorial Board of the Journal of Multimedia since 2009. He and his students received multiple paper awards including the Best Paper Awards of VCIP 2012, ICIMCS 2012, ACM MUM 2011, and the Best Student Paper Award of MobiMedia 2009.