Neural Processing Letters (2024) 56:194
https://doi.org/10.1007/s11063-024-11652-7

Temporal Diversity-Aware Micro-Video Recommendation with Long- and Short-Term Interests Modeling

Pan Gu[1] · Haiyang Hu[1] · Dongjing Wang[1] · Dongjin Yu[1] · Guandong Xu[2]

Accepted: 8 May 2024 / Published online: 3 June 2024

Abstract
Recommender systems have become indispensable for addressing information overload for micro-video services. They are used to characterize users' preferences from their historical interactions and recommend micro-videos accordingly. Existing works largely leverage the multi-modal contents of micro-videos to enhance recommendation performance. However, limited efforts have been made to understand users' complex behavior patterns, including their long- and short-term interests, as well as their temporal diversity preferences. In micro-video recommendation scenarios, users tend to have both stable long-term interests and dynamic short-term interests, and may feel tired after incessantly receiving numerous similar recommendations. In this paper, we propose a **T**emporal **D**iversity-aware micro-**video rec**ommender (TD-VideoRec) for user behavior modeling, simultaneously capturing users' long- and short-term preferences. Specifically, we first adopt a user-centric attention mechanism to cope with long-term interests. Then, we utilize an attention network on top of a long-short term memory network to obtain users' short-term interests. Finally, a temporal diversity coefficient is introduced to characterize the temporal diversity preferences of users' click behaviors. The value of the coefficient depends on the distinction between users' long- and short-term interests extracted by vector orthogonal projection. Extensive experiments on two real-world datasets demonstrate that TD-VideoRec outperforms state-of-the-art methods.

Keywords Micro-video recommendation · Long- and short-term interests · Temporal diversity preferences

1 Introduction

Micro-videos, with lengths counted in seconds, have the characteristics of strong social attributes, low creation thresholds, and fast transmission speeds. Micro-videos can be produced simply with a smartphone, resulting in a dramatic increase in their numbers. Taking

✉ Haiyang Hu
huhaiyang@hdu.edu.cn

Pan Gu
gupan@hdu.edu.cn

1 School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China

2 Advanced Analytics Institute, University of Technology Sydney, Sydney, Australia

Kuaishou[1] in China as an example, the platform recorded an average monthly upload of around 1.1 billion micro-videos and attracted 305 million daily active users in the 9 months ended September 30, 2020. Moreover, as of March 2022, each active user on average spent over 2 h per day on Kuaishou App. The massive amount of micro-videos makes it challenging for users to quickly find their desired micro-videos, thus highlighting the necessity of a recommender system.

Micro-videos have a short duration and users usually spend a relatively long time on micro-video platforms, resulting in much longer behavior sequences. Therefore, building a model for micro-video recommendation is non-trivial due to the complex behavior patterns exhibited by users:

1. **Long- and Short-Term Interests** Users typically exhibit both stable long-term interests and dynamic short-term interests. For instance, as shown in Fig. 1, a professional dancer may consistently browse dance and healthy diet videos (long-term interests), but also express interest in travel footage when planning a trip (short-term interests). Only capturing long-term interests may fail to respond to users' current dynamic demands timely. On the other side, merely focusing on short-term interests may yield homogeneous recommendations, since long-term interests commonly contain more diverse information than short-term interests. As such, it is essential to simultaneously characterize users' long- and short-term interests.
2. **Temporal Diversity Preferences** Users tend to exhibit preferences for temporal diversity on micro-video contents, that is, they prefer not to view homogeneous micro-videos successively. This differs from e-commerce recommendations, where users typically have a clear and unique shopping demand in a session [1]. It should be noticed that the temporal diversity is not exactly equivalent to diversity. Instead, it emphasizes that the recommended items tend to be diverse from the recently viewed items while still aligning with overall user preferences.
3. **Correlation Between Short-Term Interests and Temporal Diversity Preferences** Short-term interests and temporal diversity preferences represent two aspects of users' short-term preferences. Within a short timeframe, users exhibit intrinsic interests but avoid continuously viewing similar micro-videos. For example, as shown in Fig. 1, a user who has recently watched several travel videos may appreciate another recommendation related to landscapes. However, receiving numerous similar videos in succession may lead to user fatigue. In such cases, the system should prioritize long-term interests, such as dance videos. That is, excessively emphasizing a user's short-term interests will inevitably depress the temporal diversity of recommended items. Thus, it is crucial to adaptively balance the modeling of short-term interests and temporal diversity preferences.

In recent years, many effective algorithms have been proposed to characterize users' preferences for micro-video recommendation [2–12]. For example, Chen et al. [2] leveraged an item- and category-level attention network to capture user's interests, while He et al. [3] devised a hypergraph-based framework to model temporal user-item interactions and incorporated multi-modal features to enhance representation learning. Additionally, Li et al. [4] designed a shared temporal graph-based sequential framework to capture users' positive and negative preferences from multiple behaviors. These approaches attempted to introduce more auxiliary information to enhance the model performance, but none of them simultaneously captured users' long- and short-term interests or considered their temporal diversity preferences.

[1] https://www.kuaishou.com.

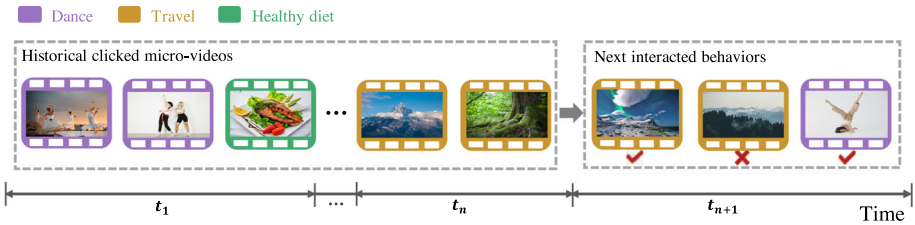


Fig. 1 Motivating examples of our TD-VideoRec model. The left part is the historical micro-videos clicked by the user, while the right part is the next interacted micro-videos

In this paper, we propose a **T**emporal **D**iversity-aware micro-**video rec**ommender, named TD-VideoRec, for better modeling users' preferences. Our TD-VideoRec model comprises three key modules: the long-term interests module, the short-term interests module, and the temporal diversity-aware module. Specifically, we devise two separate encoders to distinctively capture users' long- and short-term interests. To address long-term interests, we adopt a user-centric attention mechanism to filter out irrelevant information from users' behaviors. On the other side, we utilize an attention network atop a long-short term memory network to obtain users' short-term interests. Subsequently, a temporal diversity coefficient is introduced to characterize users' temporal diversity preferences in click behaviors. It is presumed that temporal diversity can be inferred from the distinction between users' long- and short-term interests. To extract these distinctive parts, we apply vector orthogonal projection [13] and feed them into a two-layer feed-forward network (MLP) to estimate the temporal diversity coefficient. Note that the user ID representation serves as a representative of long-term interests. Combining the outputs of the three key modules, we obtain a unified user representation. Finally, the click probability can be estimated by a two-layer MLP, which takes a concatenation of the user unified representation and the given micro-video embedding as inputs. To the best of our knowledge, we are the first to explore users' temporal diversity preferences for the micro-video recommendation. Extensive experiments on two real-world datasets demonstrate that TD-VideoRec outperforms state-of-the-art methods. The main contributions of this paper can be summarized as follows:

- We propose to highlight users' temporal diversity preferences and integrate them with long- and short-term interests for micro-video recommendation.
- We introduce a temporal diversity-aware micro-video recommender by incorporating a temporal diversity coefficient to measure the degree of users' temporal diversity preferences, which is estimated by the distinctive parts extracted from long- and short-term interests. The long- and short-term interests are captured by two separate encoders.
- We conduct extensive experiments on two real-world datasets, demonstrating the effectiveness of TD-VideoRec compared with several state-of-the-art methods.

Organization The remainder of the paper is organized as follows. In Sect. 2, we review related work. In Sect. 3, we study the effect of recently clicked micro-videos and users' temporal diversity preferences. In Sect. 4, we elaborate on the proposed TD-VideoRec model. In Sect. 5, we present the implementation details and analyze the model performance. In Sect. 6, we conclude the paper.

2 Related Work

The studies related to our research can be divided into two main paradigms: micro-video recommendation and user behavior modeling.

2.1 Micro-Video Recommendation

In recent years, many approaches based on deep learning methods have been proposed for micro-video recommendation [2–12, 14–19], including recurrent neural network (RNN) [4, 14], self-attention mechanism [2, 6, 8, 15], and graph convolution network (GCN) [3, 9, 11, 12, 17–19]. For instance, Li et al. [4] designed a temporal graph-based sequential network to capture users' positive and negative interests from multi-behaviors. For self-attention based methods, Chen et al. [2] leveraged an item- and category-level attention network to capture users' preferences. Liu and Chen [6] adopted a self-attention network to calculate the importance of multi-modal contents and then utilized a multi-head attention network to model users' interests for micro-video recommendation. Liu et al. [8] devised a user-video co-attention network to exploit multi-modal features of both users and micro-videos.

Regarding GCN-based approaches, Wei et al. [9] constructed a user-item bipartite graph in each modality and learned modal-specific representations for users and items using graph convolution network. Wei et al. [11] learned multi-level user interests from the items' co-interacted patterns with a hierarchical graph structure. He et al. [3] devised a hypergraph-based network to model the temporal user-item interactions and exploited multi-modal features to enhance the representation learning. Wei et al. [18] devised a graph-refined neural network to discover and address potential false-positive edges. Liu et al. [19] introduced a user-oriented graph denoising process to prune the noisy interactions for micro-video recommendation. More recently, Tian et al. [12] devised a sequential capsule network to transform the users' interacted sequence into multi-interests representation and further adopted a graph convolutional network to understand user preferences at a fine-grained level. Despite achieving encouraging performances, these efforts ignored users' complicated behavior patterns, thus limiting the accuracy of user representation. In addition, these methods primarily extracted users' long-term interests, which failed to respond to users' current demands timely. In this paper, we simultaneously characterize users' long- and short-term interests to facilitate the performance of recommender systems.

2.2 User Behavior Modeling

Early works [20–23] about user behavior modeling primarily relied on Markov chains and traditional matrix factorization to extract users' short- and long-term interests, respectively. Inspired by the recent success of deep learning in natural language processing, an increasing number of researchers have adopted deep learning methods in recommender systems. For example, Hidasi et al. [24] first employed recurrent neural networks to capture sequential patterns in users' behaviors, achieving impressive performance gains. Wu et al. [25] built session graphs from session sequences and further captured high-order transitions among items with graph neural networks. Moreover, attention networks [26] were introduced to capture long-range dependencies among users' historical behaviors. Particularly, Kang and McAuley [27] employed self-attention networks to model recent few actions of users for item predictions.

However, these methods mainly focus on modeling relatively recent behaviors to characterize users' preferences. Recently, a few works [1, 28–33] have been proposed to capture the long- and short-term interests simultaneously. For instance, Lv et al. [1] utilized multi-head self-attention to characterize evolving short-term interests, which were then combined with long-term interests using a gated fusion network. Zheng et al. [33] disentangled long- and short-term interests from users' historical behaviors trained under a contrastive learning framework. However, these methods simply considered users' multi-level interests, which tend to provide homogeneous and myopic recommended items. In light of this, we explicitly model users' temporal diversity preferences to get a more comprehensive user representation.

3 Motivations

To analyze how recently clicked micro-videos and temporal diversity preferences affect recommendation performance, we conduct preliminary experiments over two public datasets. The first dataset, MicroVideo-1.7M [2], is built from an anonymous micro-video platform in China. The second dataset is Kuaishou-3.0M [4], originally released by Kuaishou.[2] In these datasets, each micro-video is represented by a visual embedding of its cover picture, and each interaction consists of user ID, micro-video ID, timestamp, and whether the user clicked the micro-video. The "click" behavior means the user clicks the micro-video after previewing its thumbnail, which is collected by the micro-video platforms.

3.1 Data Analysis

For the MicroVideo-1.7M dataset, each micro-video is manually annotated with only one category out of a total of 512 categories. Each user clicks on an average of 218 micro-videos, spanning 117 categories. This suggests a highly detailed category division. We investigate whether users tend to watch micro-videos from the same category within a short period. We calculate CRF@K (Category Repetition Frequency@K) for each user's clicked sequence. CRF@K measures whether a micro-video and its K adjacently preceding micro-videos belong to the same category. To provide a baseline for comparison, we repeat the same process with randomly selected K preceding micro-videos. From the results shown in Fig. 2, we observe that adjacent micro-videos are more likely to belong to the same category compared to those further apart in the sequence. However, the repetition frequency remains relatively low. For example, only 12% of micro-videos share the same category with their adjacent 10 micro-videos. This observation suggests that users prefer to click on similar micro-videos in a short time, but also exhibit strong temporal diversity preferences.

3.2 Preliminary Experiments and Performance Comparison

We conduct three experiments employing a long-short term memory network (LSTM) to model users' behavior sequence. AUC (Area Under Curve) is adopted as our performance metric, similar to [4]. The first experiment aims to explore the effect of recently clicked micro-videos. We employ the LSTM encoder to model users' recent K micro-video clicks, intending to capture short-term interests. As shown in Fig. 3, the results indicate that the predictive capability of short-term interests is inferior to that of long-term interests but still shows

[2] https://www.kuaishou.com/activity/uimc.

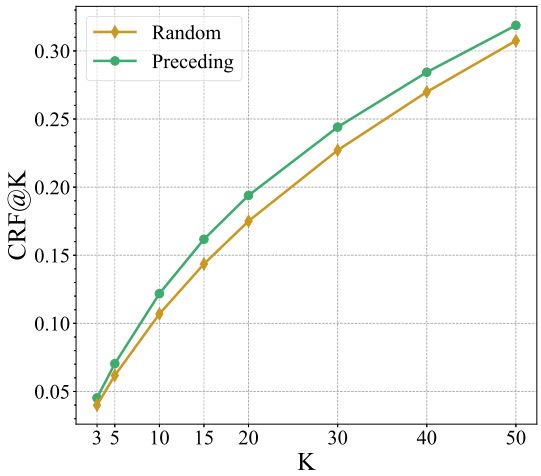
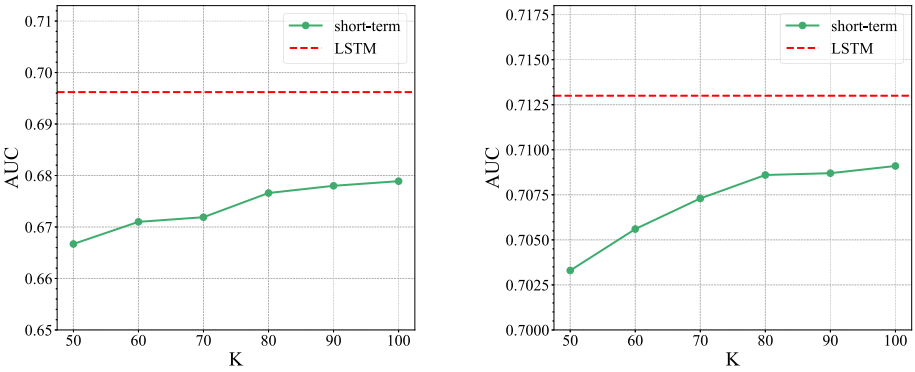


Fig. 2 Category Repetition Frequency with different numbers of preceding micro-videos



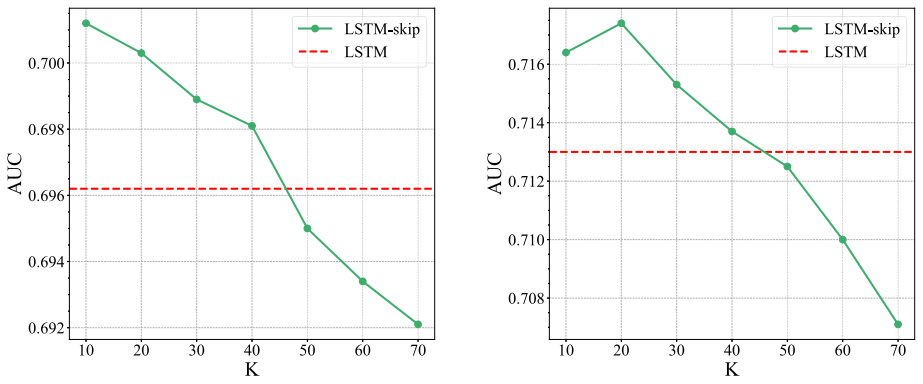
(a) MicroVideo-1.7M (b) Kuaishou-3.0M

Fig. 3 Model performance with capturing users' short-term interests

comparable performance. Notably, substantial improvements are observed as K increases from 50 to 100. Note that the entire sequence length is set to 300. Therefore, it is reasonable to conclude that although users tend to have stable interests on micro-video platforms, their short-term interests can contribute to a better understanding of their preferences.

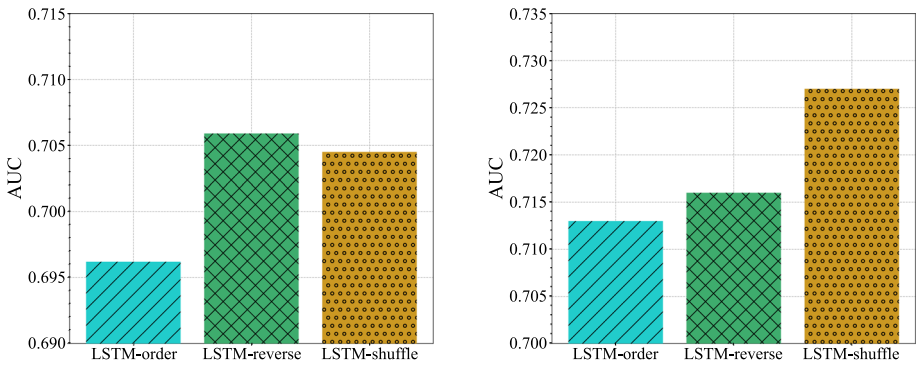
Subsequently, we conduct an experiment where the latest K clicked micro-videos are removed from the entire sequences, and the remaining ones were taken as input to train the LSTM network, namely LSTM-skip. Figure 4 illustrates the model performance with different numbers of skipped micro-videos. Interestingly, we find that removing a few clicked micro-videos actually improves the performance of the LSTM network. However, continuously removing more recent behaviors inevitably hurts the predictive capabilities of the model. This observation aligns with the reasonable assumption that though recent behaviors provide valuable information for short-term interests, modeling the most recently clicked micro-videos may adversely affect users' preferences for temporal diversity.

Finally, we utilize the vanilla LSTM method to model users' interests, but this time we inverse and randomly shuffle the originally ordered sequences separately, named LSTM-reverse and LSTM-shuffle. We also train LSTM-order, where the sequences remain in



(a) MicroVideo-1.7M (b) Kuaishou-3.0M

Fig. 4 Model performance w.r.t. the different numbers of skipped micro-videos



(a) MicroVideo-1.7M (b) Kuaishou-3.0M

Fig. 5 Performance comparison with various data processing strategies

their original order. As seen in Fig. 5, both LSTM-reverse and LSTM-shuffle outperform the LSTM-order model. The results further demonstrate that overly emphasizing users' short-term interests may hurt the temporal diversity of recommended micro-videos, leading to suboptimal performance. Moreover, LSTM-shuffle achieves obvious improvement over LSTM-order, especially in the Kuaishou-3.0M dataset. One possible explanation could be that, through randomly shuffled sequences, the LSTM model can better capture the long-term dependencies between micro-videos.

3.3 Conclusion and Motivations

Based on these statistical results, we can conclude that users exhibit both long- and short-term interests, as well as temporal diversity preferences simultaneously. However, excessively focusing on the users' short-term interests may hurt the temporal diversity of recommended micro-videos, yielding suboptimal performance. Therefore, we propose a model that considers both users' long- and short-term interests, along with temporal diversity preferences.

Furthermore, we adaptively adjust the weight of short-term interests to balance short-term interests and temporal diversity preferences.

4 Proposed Model

We first formulate the micro-video recommendation problem and then introduce our proposed TD-VideoRec model by describing its main modules in detail.

4.1 Notations and Problem Definition

Let $\mathcal{U}$ and $\mathcal{V}$ denote the sets of users and items, respectively. $\mathcal{B}(u) = \{v_1^u, v_2^u, v_3^u, \ldots, v_n^u\}$ represents the click behaviors of user $u \in \mathcal{U}$ in chronological order, where n is the sequence length and v_j^u is the j-th item clicked by user u. The long-term behavior sequence of user u is denoted by $\mathcal{L}^u$, which is exactly the $\mathcal{B}(u)$. The recently clicked micro-videos of user u are regarded as the short-term behavior sequence, namely $\mathcal{S}^u = \{v_{n-K+1}^u, v_{n-K+2}^u, v_{n-K+3}^u, \ldots, v_n^u\}$, where K is the sequence length. For simplicity, we omit the superscript u of items in the following. Based on these mathematical notations, the micro-video recommendation problem is formulated as follows:

- **Input:** The long-term behavior sequence $\mathcal{L}^u$ and short-term behavior sequence $\mathcal{S}^u$ of user u , as well as the given micro-video v_{new}.
- **Output:** The click probability of user u on the micro-video v_{new}.

The mathematical notations are summarized in Table 1. Additionally, $\mathbf{x}_j \in \mathbb{R}^d$ is the visual embedding of micro-video v_j, where d is the vector length. It is reasonable to represent each micro-video by its visual embedding of the cover picture. Micro-videos are typically short in duration, and the thumbnail is often the most representative snapshot. Moreover, cover pictures provide users with first impressions, which significantly influence their decision to click on micro-videos [2, 16].

Table 1 Notations summary

Notations	Descriptions
u	A user
v	A micro-video
v_{new}	A given micro-video
$\mathcal{B}(u)$	A user' clicked sequence
$\mathcal{L}^u$	The long-term behaviors of user u
$\mathcal{S}^u$	The short-term behaviors of user u
K	The number of short-term behaviors
$\mathbf{x}_j$	The visual embedding of micro-video v_j
$\mathbf{m}^u$	The embedding of user u
$\mathbf{p}_l^u$	The long-term interests of user u
$\mathbf{p}_s^u$	The short-term interests of user u
γ^u	The temporal diversity coefficient of user u
$\mathbf{p}^u$	The final user representation of user u

4.2 User Preference Modeling

To obtain a comprehensive user representation for micro-video recommendation, we propose to jointly model users' long- and short-term interests and temporal diversity preferences. Users' long-term interests represent their static and inherent preferences, while short-term interests involve their dynamic and time-evolving preferences. Motivated by the distinct nature, recent studies [1, 32, 33] design different encoders to learn long- and short-term interests separately. We also discriminatively learn the two aspects of users' interests with different encoders, aligning with the insights gained from these recent work. Moreover, users on micro-video platforms exhibit temporal diversity preferences, encouraging the recommendation system to recommend micro-videos different from those recently clicked. We have to mention that the temporal diversity, as discussed in our work, aligns with the concept represented in the work [34]. However, it slightly differs from the definition of novelty defined in the work [35], where novelty is defined in terms of items that have not been previously recommended to a user. Although Wu et al. [34] introduce the concept of temporal diversity, they do not provide a formal formulation for it, nor do they evaluate recommendation methods in terms of temporal diversity metric. In our paper, we define a new evaluation metric **TD@N**(Temporal Diversity@N) to measure the temporal diversity of recommended results.

To achieve the aforementioned goals, we propose a temporal diversity-aware micro-video recommendation method to obtain a unified user representation, which can be formulated as follows:

$$\begin{cases} \mathbf{p}^u = \mathbf{p}_l^u + \mathbf{p}_s^u - \gamma^u \mathbf{p}_s^u & (1) \\ \mathbf{p}_l^u = f_1(\mathcal{L}^u, \mathcal{U} \mid \theta_1) & (2) \\ \mathbf{p}_s^u = f_2(\mathcal{S}^u \mid \theta_2) & (3) \\ \gamma^u = f_3(\mathbf{p}_l^u, \mathbf{p}_s^u, \mathcal{U} \mid \theta_3) & (4) \end{cases}$$

where f_1, f_2 and f_3 are the trainable functions to learn long-term interests $\mathbf{p}_l^u$, short-term interests $\mathbf{p}_s^u$ and temporal diversity coefficient γ^u of user u. θ_1, θ_2, and θ_2 are trainable model parameters. $\mathcal{U}$ denotes the user profiles and are exactly the user IDs in our scenario. The user preference model ζ captures complicated behavior patterns through three modules: the long-term interests module f_1, the short-term interests module f_2, and the temporal diversity-aware module f_3. The overall architecture is depicted in Fig. 6. We now describe these three modules in detail.

- **Long-Term Interests Module in Eq. (2).** Long-term interests are usually luxuriant and change slowly over time, which can be captured from historical behaviors and user profiles. Therefore, we take $\mathcal{L}^u$ and $\mathcal{U}$ as inputs to train f_1.
- **Short-Term Interests Module in Eq. (3).** Short-term interests represent users' current motivations, which are dynamic and evolving relatively. We hence train f_2 based on short-term behavior sequence $\mathcal{S}^u$.
- **Temporal Diversity-Aware Module in Eq. (4).** Users would feel less satisfied if the system keeps recommending micro-videos that appeal to the user's current motivations. To model temporal diversity preferences, we design a temporal diversity coefficient γ^u to reduce the weight of short-term interests. The value of diversity coefficient γ^u is relevant to the distinction between long- and short-term interests, we hence include $\mathbf{p}_l^u$, $\mathbf{p}_s^u$ and $\mathcal{U}$ as the inputs of f_3.

Note that short-term interests and temporal diversity preferences are two aspects of users' short-term preferences. Over a short period, users exhibit intrinsic interests but are unwilling to continuously click similar micro-videos. Intuitively, long-term interests express more comprehensive and diversified information than short-term interests, and those who have diversified long-term interests tend to have stronger preferences for temporal diversity on items. Therefore, the temporal diversity coefficient depends on the distinction between long- and short-term interests.

4.3 Long-Term Interests Module

In this module, we devise a user-centric attention network to extract users' long-term interests, as shown in Fig. 6a. Specifically, we utilize an attention network with profile information as the query to extract the long-term interests $\mathbf{p}_l^u$ of user u from the long-term behavior sequence $\mathcal{L}^u$. Formally, the process can be formulated as:

$$\mathbf{p}_l^u = \sum_{j=1}^{n} \alpha_j \cdot \mathbf{x}_j \tag{5}$$

$$\alpha_j = \mathbf{x}_j^\top \mathbf{W}_l \mathbf{m}^u \tag{6}$$

$$\alpha_j = \frac{exp(\alpha_j)}{\sum_{\tau=1}^{n} exp(\alpha_\tau)} \tag{7}$$

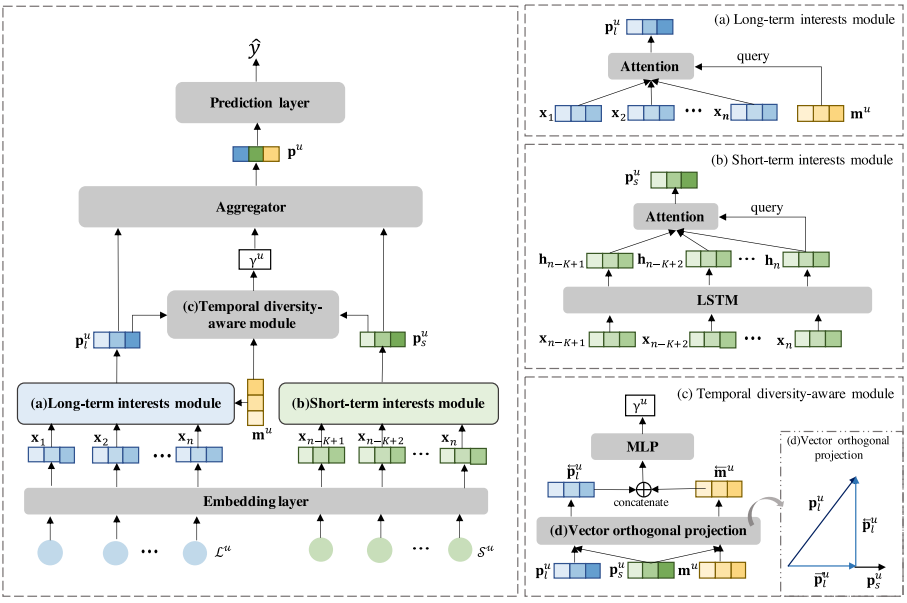


Fig. 6 The overview of the proposed model TD-VideoRec. The bottom left part illustrates the overall model schema. The right three parts illustrate the core model modules, including **a** the long-term interests module, **b** the short-term interests module, and **c** the temporal diversity-aware module. Besides, **d** describes the working of the vector orthogonal projection process

where $\mathbf{W}_l \in \mathbb{R}^{d\times d}$ is a transformation matrix. $\mathbf{m}^u$ is the representation of the user ID, obtained by a lookup operation from user embedding matrix $\mathbf{M}$. α_j reflects the attention coefficient of j-th micro-video from the long-term behavior sequence $\mathcal{L}^u$, capturing the correlations between global information and the i-th micro-video.

4.4 Short-Term Interests Module

When it comes to short-term interests, recurrent neural network-based methods have shown powerful performance [24]. Concretely, we apply an attention network on top of a long-short term memory network (LSTM) to characterize users' short-term interests, as illustrated in Fig. 6b. We first utilize an LSTM network to generate hidden vectors, and the last hidden vector is used as the query vector. Second, an attentive network is employed to extract short-term interests $\mathbf{p}_s^u$ from the hidden vectors. Formally, the process can be formulated as follows:

$$\{\mathbf{h}_{n-K+1}, \mathbf{h}_{n-K+2}, \cdots, \mathbf{h}_n\} = LSTM(\{\mathbf{x}_{n-K+1}, \mathbf{x}_{n-K+2}, \ldots, \mathbf{x}_n\}) \tag{8}$$

$$\mathbf{p}_s^u = \sum_{i=n-K+1}^{n} \beta_i \cdot \mathbf{h}_i \tag{9}$$

$$\beta_i = \mathbf{h}_i^\top \mathbf{W}_s \mathbf{h}_n \tag{10}$$

$$\beta_i = \frac{exp(\beta_i)}{\sum_{\tau=n-K+1}^{n} exp(\beta_\tau)} \tag{11}$$

where $\mathbf{W}_s \in \mathbb{R}^{d\times d}$ is a transformation matrix, and β_i denotes the attention value. $\{\mathbf{x}_{n-K+1}, \mathbf{x}_{n-K+2}, \ldots, \mathbf{x}_n\}$ is the embedding of the short-term behaviors $\mathcal{S}^u$. By aggregating the long- and short-term interests, we can generate the user representation $\mathbf{p}^u = \mathbf{p}_l^u + \mathbf{p}_s^u$.

4.5 Temporal Diversity-Aware Module

Excessive emphasis on short-term interests can lead to homogeneous recommendations, which harms the users' preferences for the temporal diversity of micro-video content, especially for those with diverse interests. Previous work [34] modeled users' temporal diversity characteristics by learning a shared weight parameter to combine their global and recent interests, failing to capture the variations in users' preferences for temporal diversity.

We argue that users' temporal diversity preferences can be depicted by the distinction between their long- and short-term interests. Long-term interests encapsulate users' overall preferences, while short-term interests merely indicate their current preferences. Therefore, we decompose users' long-term interests into two parts: those relevant to short-term interests (common part) and those distinct from short-term interests (distinctive part). As illustrated in Fig. 6c, we first apply vector orthogonal projection [13] to extract the distinctive part from two groups of pairs $\langle \mathbf{p}_l^u, \mathbf{p}_s^u \rangle$ and $\langle \mathbf{m}^u, \mathbf{p}_s^u \rangle$, where $\mathbf{p}_l^u$ and $\mathbf{m}^u$ are all representatives of long-term interests. Subsequently, we feed them into a two-layer feed-forward network (MLP) to achieve the temporal diversity coefficient γ^u. Figure 6d illustrates the orthogonal projection process of the pair $\langle \mathbf{p}_l^u, \mathbf{p}_s^u \rangle$ in a two-dimensional space, which evolves in two steps as follows:

$$\overrightarrow{\mathbf{p}}_l^u = project(\mathbf{p}_l^u, \mathbf{p}_s^u) = \frac{\mathbf{p}_l^u \cdot \mathbf{p}_s^u}{|\mathbf{p}_s^u|} \frac{\mathbf{p}_s^u}{|\mathbf{p}_s^u|} \tag{12}$$

$$\overleftarrow{\mathbf{p}}_l^u = \mathbf{p}_l^u - \overrightarrow{\mathbf{p}}_l^u \tag{13}$$

Algorithm 1: TD-VideoRec Algorithm

Input: user sets $\mathcal{U}$, item embedding matrix X, historical behaviors $\mathcal{B}(u)$, learning rate η, the length of recent behaviors K
Output: model parameters Θ

```
Initialize Θ from Normal Distribution N(0, 0.01) ;
repeat
    generate the set of observations {(L^u, S^u, v_new)};
    for each observation {(L^u, S^u, v_new)} do
        capture p_l^u according to Eqs. (5)–(7) ;
        capture p_s^u according to Eqs. (8)–(11) ;
        compute γ^u according to Eqs. (12)–(14);
        get p^u according to Eq. (1) ;
        compute ŷ according to Eqs. (15)–(16);
        update Θ with Adam optimizer ;
    end
until Convergence;
return Θ
```

where $project()$ denotes the projection function. The vector $\overleftarrow{\mathbf{p}}_l^u$ represents the distinction between $\mathbf{p}_l^u$ and $\mathbf{p}_s^u$. Likewise, the vector $\overleftarrow{\mathbf{m}}^u$ is obtained for the pair $\langle \mathbf{m}^u, \mathbf{p}_s^u \rangle$. Based on the distinction vectors $\overleftarrow{\mathbf{p}}_l^u$ and $\overleftarrow{\mathbf{m}}^u$, the user-specific diversity coefficient γ^u can be defined as follows:

$$\gamma^u = \mathbf{W}_2 \phi(\mathbf{W}_1 \left[\overleftarrow{\mathbf{p}}_l^u, \overleftarrow{\mathbf{m}}^u \right] + \mathbf{b}_1) + b_2 \tag{14}$$

where $\left[\overleftarrow{\mathbf{p}}_l^u, \overleftarrow{\mathbf{m}}^u \right]$ represents a concatenation of vectors. $\mathbf{W}_1 \in \mathbb{R}^{d \times 2d}$ and $\mathbf{W}_2 \in \mathbb{R}^{1 \times d}$ are transform matrices, $\mathbf{b}_1 \in \mathbb{R}^d$ and $b_2 \in \mathbb{R}$ are a bias vector and bias scalar. ϕ is the RELU activation function.

The final user representation $\mathbf{p}_u$ is calculated by combining $\mathbf{p}_l^u$, $\mathbf{p}_s^u$, and γ^u according to Eq. (1). Equation (1) can be understood as a way to strike a balance between modeling short-term interests and accounting for temporal diversity preferences. Through the temporal diversity coefficient, the weight of short-term interests can be reduced when necessary. Furthermore, Eq. (1) can also be regarded as a fusion method to adaptively merge the long- and short-term interests according to the users' temporal diversity characteristics.

4.6 Prediction Layer

In this section, the user's click probability towards the candidate micro-video v_{new} is predicted. Similar to [2], the user representation $\mathbf{p}^u$ and the representation of the given micro-video v_{new} are passed to a two-layer MLP as follows:

$$\mathbf{f}^u = \phi(\mathbf{W}_f \left[\mathbf{p}^u, \mathbf{x}_{new} \right] + \mathbf{b}_f) \tag{15}$$

$$\hat{y} = \mathbf{W}_y \mathbf{f}^u + b_y \tag{16}$$

where $\mathbf{W}_f \in \mathbb{R}^{2d \times 2d}$ and $\mathbf{W}_y \in \mathbb{R}^{1 \times 2d}$ are transformation matrices, $\mathbf{b}_f \in \mathbb{R}^{2d}$ and $b_y \in \mathbb{R}$ are a bias vector and bias scalar, and ϕ denotes the ReLU activation function. $\mathbf{x}_{new}$ is the representation of the micro-video v_{new}. The model minimizes the sigmoid cross-entropy loss function as follows:

$$L(y, \hat{y}) = -(y \log \sigma(\hat{y}) + (1 - y) \log(1 - \sigma(\hat{y})) \tag{17}$$

where $y \in \{0, 1\}$ is the ground truth and σ is the *sigmoid* activation function. The complete process is detailed in Algorithm 1.

5 Experiments

5.1 Datasets

MicroVideo-1.7M The dataset is built from an anonymous micro-video platform in China [2]. It contains 10,986 users and their 12,737,619 interactions among 1,704,880 micro-videos. Each Micro-video is manually annotated with only one category and represented by a 512-d vector extracted from its cover picture by the Inception-v3 model. Interactions are divided into training and testing sets based on micro-videos.

Kuaishou-3.0M Originally released by Kuaishou, this dataset is a public version constructed by [4]. We further filter out micro-videos clicked by more than 200 users. Consequently, 10,000 users are selected, and their 10,692,289 interactions among 3,239,534 micro-videos are included. Each micro-video is represented as a 2048-d visual embedding of its cover picture. For experimental analysis, we split users' first 80% of behaviors for training and the remaining for testing.

Table 2 summarizes the basic statistics of the two datasets. In the MicroVideo-1.7M dataset, micro-videos in the test set are newly generated, specially designed for cold-start recommendation. While for the Kuaishou-3.0M dataset, 42% of micro-videos in the test set have appeared in the training set. Each behavior in these two datasets is associated with a user ID, item ID, timestamp, and whether the user clicked the micro-video. Here, "click" behavior means the user clicks the micro-video after previewing its cover picture. The timestamps have been processed to preserve only the sequential order, while absolute time information is unknown. Before training our proposed model, the dimension of each micro-video embedding is reduced to 64-d using the principal component analysis method, to reduce the model complexity. Furthermore, we split the training sets into two subsets for training and validation, where the ratio is 9:1.

5.2 Evaluation Metrics

The micro-video recommendation is formulated as a binary classification problem. Therefore, we adopt the Area Under Curve (AUC) as the primary evaluation metric, which measures the

Table 2 Basic statistics of the datasets after preprocessing

Statistics	MicroVideo-1.7M	Kuaishou-3.0M
# users	10,986	10,000
# micro-videos	1,704,880	3,231,539
# total interactions	12,737,619	10,692,289
# interactions per user	1159	1069
# clicked interactions per user	218	201
# train interactions	8,970,310	8,507,916
# test interactions	3,767,309	2,184,373

model's ability to distinguish between positive and negative instances. In addition, we adopt other metrics, including P@N, R@N, and F1@N, where N is set to 50, in line with previous works [2, 4, 10, 16]. Given a user's recommendation list, P@N denotes the proportion of interested items in the top-N recommended items, R@N reflects the coverage ability of top-N recommended items, and F1@N is derived by a combination of precision and recall.

5.3 Baselines

The following methods are used to compare with our proposed model.

- **BPR** [36] BPR is a widely used method for non-sequential recommendation, which optimizes a pairwise ranking objective function.
- **NCF** [37] Neural collaborative filtering learns latent vectors for users and items with a deep neural architecture.
- **ATRank** [38] ATRank uses self-attention with time encoding to model user behaviors for item recommendation.
- **LSTM** [39] Long-short term memory network is a classical sequential model widely used for item recommendation.
- **THACIL** [2] THACIL first splits user behaviors into multiple blocks and builds block representation with an item- and category-level attention. Then a multi-head self-attention is adopted to characterize users' interests from block sequences.
- **ALPINE** [4] ALPINE is a temporal graph-based recommendation system designed to capture users' interested and uninterested information from clicked and unclicked samples simultaneously.
- **CLSR** [33] CLSR disentangles long- and short-term interests from users' behavior sequences trained under a contrastive learning framework.
- **TempRec** [34] TempRec is a temporal diversity-aware news recommendation framework, compatible with any sequence encoders. We implement this model with the same encoders as our model. The final click probability of a given item is estimated by a learnable weighted summation of users' long- and short-term interests.

5.4 Implementation Details

For all the aforementioned methods, a user is represented by a 64-d vector randomly initialized with a Gaussian distribution, while a micro-video is embedded into a 64-d visual vector (except for THACIL). In the THACIL model, each micro-video is embedded into a 128-d vector, with 64-d category embedding and 64-d visual information embedding. The lengths of the long-term and short-term behavior sequences are both set to 300. If the sequence exceeds 300, we only preserve as many as 300 micro-videos. For training, Adam optimizer [40] is adopted with a learning rate of 0.001, where the batch size is set as 1024. The model is trained in TensorFlow [41] on a GeForce GTX 2080 Ti GPU. The source code is available at https://github.com/gupanz/TD_VideoRec.

5.5 Overall Performance Comparison

We conduct experiments over the two datasets to demonstrate the effectiveness of TD-VideoRec compared to state-of-the-art recommendation methods. As shown in Table 3, we have the following observations:

Table 3 Overall performance comparison of all methods

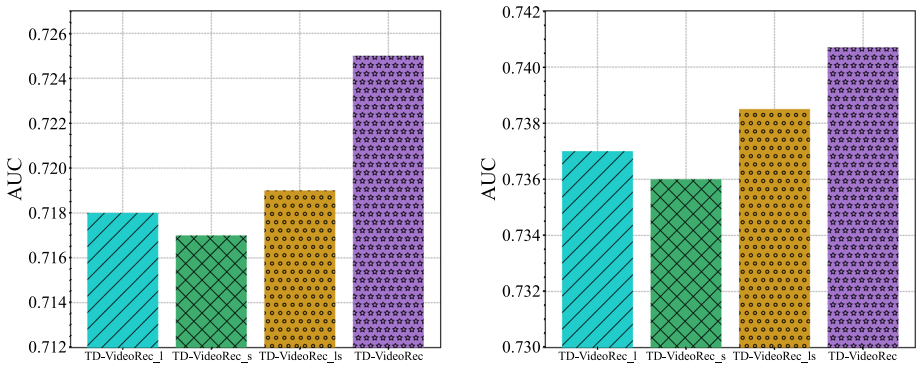
Methods	MicroVideo-1.7M				Kuaishou-3.0M			
	AUC	P@50	R@50	F1@50	AUC	P@50	R@50	F1@50
BPR	0.620	0.236	0.395	0.295	0.614	0.274	0.480	0.348
NCF	0.702	0.282	0.442	0.344	0.723	0.292	0.516	0.373
LSTM	0.696	0.269	0.447	0.336	0.713	0.293	0.495	0.368
ATRank	0.698	0.295	0.461	0.360	0.717	0.294	0.519	0.376
THACIL	0.667	0.303	0.465	0.367	0.723	0.302	0.531	0.385
ALPINE	0.713	0.300	0.460	0.362	0.733	0.300	0.530	0.383
CLSR	0.710	0.296	0.456	0.359	0.729	0.299	0.527	0.382
TD-VideoRec	**0.725**†	**0.308**	**0.468**	**0.372**	**0.741**†	**0.307**	**0.535**	**0.390**
TempRec	0.716	0.302	0.462	0.365	0.738	0.304	0.533	0.387

The TempRec model is implemented with the same encoders as ours. The best and the second best results are highlighted in boldface and underline. †indicates the TD-VideoRec surpasses the best baseline in terms of AUC at p-value < 0.01

- BPR performs worst for micro-video recommendation because it does not extract users' interests from their behavior sequences. Therefore, it fails to capture the transition between items and the users' evolving interests. Though NCF is not a sequential model, it still shows competitive performance. This suggests that the deep network, a concatenation of the user and item vector followed by several MLP layers, can better capture the correlation between users and items.
- Sequential methods, such as LSTM, ATRank, THACIL, and ALPINE, outperform the BPR model. This further indicates that modeling users' behaviors as a sequence is essential. Moreover, attention-based methods, including ATRank, THACIL, and ALPINE surpass LSTM networks. This signifies that the attention network is effective in capturing global interests over a long sequence by focusing on the core information. ALPINE further improves recommendation quality by linking two micro-videos that share similar visual information.
- CLSR and TempRec achieve promising performance compared with the other baselines. This verifies the necessity of modeling long- and short-term interests simultaneously. Especially the TempRec model, implemented by the same encoders in our model, achieves good results. This partly demonstrates the effectiveness of our TD-VideoRec model.
- TD-VideoRec achieves the best performance, which signifies that our model can get a better user representation for micro-video recommendation. Particularly, our TD-VideoRec model surpasses TempRec, indicating the efficacy of the temporal diversity-aware module. It's worth noting that improving metrics such as precision, recall, and F1 is challenging, as also observed in other works [16, 42]. This challenge primarily arises from the significant class imbalance between positive and negative instances.

5.6 Model Analysis

Model Modules Study We study the variants of our model to further investigate the effectiveness of the long-term interests module, short-term interests module, and temporal diversity-ware module.



(a) MicroVideo-1.7M (b) Kuaishou-3.0M

Fig. 7 Ablation study of TD-VideoRec on two datasets

- **TD-VideoRec_l** For a user, only the long-term interests are considered to compute the click probability.
- **TD-VideoRec_s** For a user, merely the short-term interests are utilized to compute the click probability.
- **TD-VideoRec_ls** For a user, a simple summation of the long- and short-term interests is used to compute the click probability.

As shown in Fig. 7, TD-VideoRec_ls achieves better performance compared with TD-VideoRec_l and TD-VideoRec_s on both two datasets. This indicates the necessity to consider users' long- and short-term interests simultaneously. Moreover, TD-VideoRec outperforms TD-VideoRec_l, TD-VideoRec_s and TD-VideoRec_ls consistently on two datasets. This supports our motivation that users in micro-video recommendation scenarios show strong preferences for temporal diversity. On the other side, it also validates the effectiveness of the temporal diversity-aware module.

Temporal Diversity of recommended Items We define a new evaluation metric, called **TD@N**(Temporal Diversity@N), to measure the temporal diversity of recommended results. TD@N calculates the category difference between top-N recommended items and recently clicked N items in training sets. More formally, let $\hat{I}_{u,N}$ denotes the category sets of top-N recommended items for user u and $I_{u,N}$ denotes the category sets of recently clicked N items for user u, the category difference of the two sets denotes the members of $\hat{I}_{u,N}$ that are not in $I_{u,N}$:

$$TD@N = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{\left| \hat{I}_{u,N} \backslash I_{u,N} \right|}{N} \tag{18}$$

Figure 8 visualizes the performance on the MicroVideo-1.7M dataset. The analysis of the Kuaishou-3.0M dataset is omitted as the category attributes of micro-videos are not released. Note that the TRUTH denotes the actually clicked items in the test sets. It can be observed that (1) TD@N decreases when N ranges from 10 to 50. (2) The results recommended by TD-VideoRec_l is more temporally diverse than TD-VideoRec_s. Besides, the TD@N metric of TD-VideoRec is between TD-VideoRec_l and TD-VideoRec_s, since TD-VideoRec is an adaptive combination of TD-VideoRec_l and TD-VideoRec_s. (3)

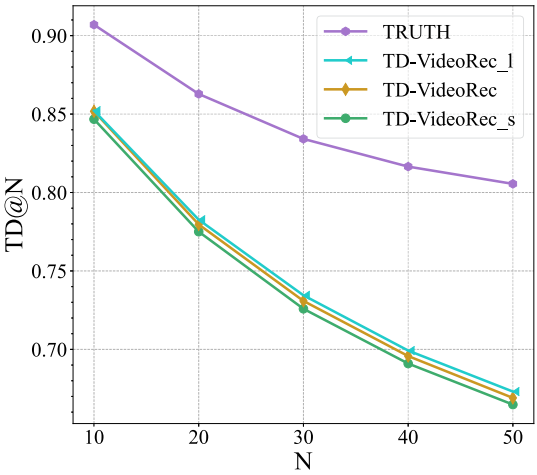


Fig. 8 Recommendation performance in terms of Temporal Diversity@N metric

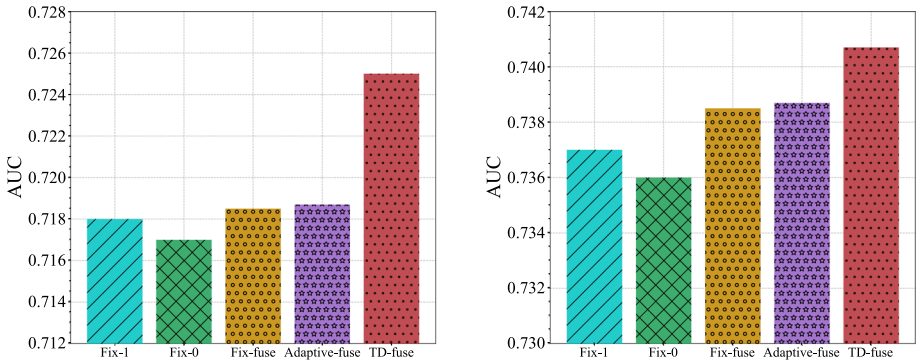
Recommendation models tend to inadvertently narrow users' interests while attempting to capture users' interests from historical interactions.

Temporal Diversity-Aware Fusion The temporal diversity-aware module can be regarded as a kind of adaptive fusion, abbreviated as TD-fuse. To verify the effectiveness of the TD-fuse, we compare it with a fixed summation of long- and short-term interests and a widely adopted adaptive fusion function, named Adaptive-fuse. Adaptive-fuse, proposed in SLi-Rec [32], is an attention-based adaptive fusion which can be formulated by $\alpha = \sigma(W_m \left[\mathbf{p}_l^u, \mathbf{p}_s^u\right] + b_m)$, $\mathbf{p}^u = \alpha \cdot \mathbf{p}_l^u + (1 - \alpha) \cdot \mathbf{p}_s^u$, where $W_m \in \mathbb{R}^{1\times 2d}$ is transformation matrix, $b_m \in \mathbb{R}$ is a bias scalar. Figure 9 demonstrates the results in terms of the AUC metric. Fix-1 and Fix-0 indicate the variants where merely a long-term or short-term interests module is used. Fix-fuse denotes a fixed optimal value obtained by grid search. We can find that our proposed TD-fuse significantly outperforms Fix-fuse and Adaptive-fuse, which suggests the effectiveness of TD-fusion of long- and short-term interests.

Influent of Recent Behaviors Length In TD-VideoRec model, we capture users' short-term interests from recently clicked micro-videos. Figure 10 illustrates the AUC results of two datasets with short-term behavior lengths K ranging from 60 to 420. The experiments on two datasets exhibit similar results. Specifically, the model performs best when $K = 300$ in the Kuaishou-3.0M dataset and $K = 360$ in the MicroVideo-1.7M dataset, respectively. When K exceeds 240, there is a marginal change in model performance for the MicroVideo-1.7M dataset. Therefore, We uniformly set $K = 300$ in our paper for two datasets. Besides, when K becomes either too large or too small, the performance of the model deteriorates. This is because a small K value fails to fully characterize short-term interests, while a large K introduces more noise.

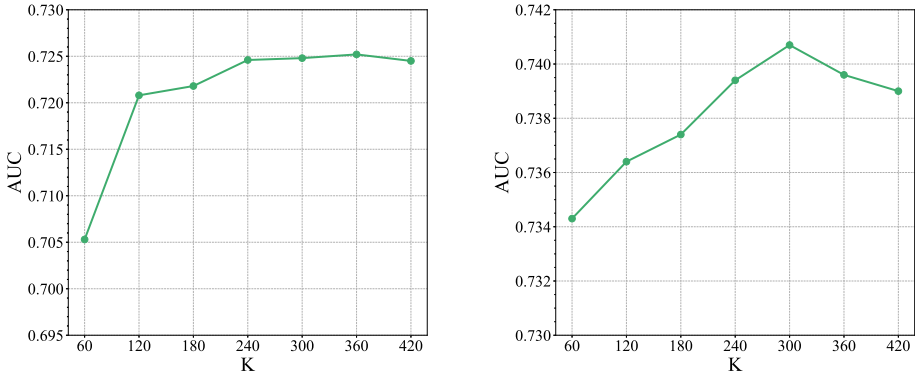
6 Conclusion and Future Work

In this paper, we first conduct data analysis on large-scale real-world datasets to illustrate the effect of users' most recently interacted behaviors. The results show that while short-term behaviors help to characterize user preferences, overly emphasizing them can



(a) MicroVideo-1.7M (b) Kuaishou-3.0M

Fig. 9 Performance comparison with different fusing methods



(a) MicroVideo-1.7M (b) Kuaishou-3.0M

Fig. 10 Recommendation performance with different short-term behaviors length

lead to suboptimal performance due to users' preference for temporal diversity. Motivated by these observations, we propose a Temporal Diversity-aware micro-video recommender (TD-VideoRec) to integrate the long- and short-term interests, as well as temporal diversity preferences for better user modeling. Specifically, we devise two separate encoders to distinctively characterize users' static long-term interests and dynamic short-term interests. To describe users' temporal diversity preferences, we define a temporal diversity coefficient which can be depicted by the distinction between users' long- and short-term interests. In addition, we devise a new evaluation metric called TD@N to measure the temporal diversity of recommended results. Extensive experiments on two public micro-video datasets demonstrate that our proposed TD-VideoRec model significantly outperforms the state-of-the-art baselines.

In future work, we plan to incorporate multi-type interactions into the interest encoders and temporal diversity-aware module, such as "like", "follow" and "comment" behaviors. This will allow us to gain deeper insights into users' diversity preferences on multiple types of interactions. Additionally, users engage with different behaviors in chronological order and the "click" action is affected by the previous "unclick" behaviors. How to model each

user's interactive behavior sequence in a holistic way is another important future work. Moreover, only the cover picture is used to represent micro-video in our work. Incorporating more comprehensive video representations, such as frame-level features, can contribute to model performance. We plan to consider watch time attributes and the frame-level features to obtain more fine-grained user preferences based on recently released frame-level micro-video datasets [43].

Acknowledgements This work was supported by the Zhejiang Provincial Key Science and Technology "LingYan" Project Foundation (2023C01145).

References

1. Lv F, Jin T, Yu C, Sun F, Lin Q, Yang K, Ng W (2019) Sdm: sequential deep matching model for online large-scale recommender system. In: Proceedings of the 28th ACM international conference on information and knowledge management, pp 2635–2643
2. Chen X, Liu D, Zha Z-J, Zhou W, Xiong Z, Li Y (2018) Temporal hierarchical attention at category- and item-level for micro-video click-through prediction. In: Proceedings of the 26th ACM international conference on multimedia, pp 1146–1153
3. He L, Chen H, Wang D, Jameel S, Yu P, Xu G (2021) Click-through rate prediction with multi-modal hypergraphs. In: Proceedings of the 30th ACM international conference on information & knowledge management, pp 690–699
4. Li Y, Liu M, Yin J, Cui C, Xu X-S, Nie L (2019) Routing micro-videos via a temporal graph-guided recommendation system. In: Proceedings of the 27th ACM international conference on multimedia, pp 1464–1472
5. Huang L, Luo B (2017) Personalized micro-video recommendation via hierarchical user interest modeling. In: Pacific Rim conference on multimedia. Springer, pp 564–574
6. Liu S, Chen, Z (2019) Sequential behavior modeling for next micro-video recommendation with collaborative transformer. In: 2019 IEEE international conference on multimedia and expo (ICME). IEEE, pp 460–465
7. Ma J, Wen J, Zhong M, Chen W, Zhou X, Indulska J (2019) Multi-source multi-net micro-video recommendation with hidden item category discovery. In: International conference on database systems for advanced applications. Springer, pp 384–400
8. Liu S, Chen Z, Liu H, Hu X (2019) User-video co-attention network for personalized micro-video recommendation. In: The World Wide Web conference, pp 3020–3026
9. Wei Y, Wang X, Nie L, He X, Hong R, Chua T-S (2019) Mmgcn: multi-modal graph convolution network for personalized recommendation of micro-video. In: Proceedings of the 27th ACM international conference on multimedia, pp 1437–1445
10. Jiang H, Wang W, Wei Y, Gao Z, Wang Y, Nie L (2020) What aspect do you like: multi-scale time-aware user interest modeling for micro-video recommendation. In: Proceedings of the 28th ACM international conference on multimedia, pp 3487–3495
11. Wei Y, Wang X, He X, Nie L, Rui Y, Chua T-S (2021) Hierarchical user intent graph network for multimedia recommendation. IEEE Trans Multimed 24:2701–2712
12. Tian Y, Chang J, Niu Y, Song Y, Li C (2022) When multi-level meets multi-interest: a multi-grained neural model for sequential recommendation. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp 1632–1641
13. Qin Q, Hu W, Liu B (2020) Feature projection for improved text classification. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics, pp 8161–8171

14. Gu P, Hu H (2024) A holistic view on positive and negative implicit feedback for micro-video recommendation. Knowl Based Syst 284:111299
15. Yu Y, Jin B, Song J, Li B, Zheng Y, Zhuo W (2022) Improving micro-video recommendation by controlling position bias. In: Machine learning and knowledge discovery in databases—European conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part I, pp 508–523
16. Han Y, Gu P, Gao W, Xu G, Wu J (2021) Aspect-level sentiment capsule network for micro-video click-through rate prediction. World Wide Web 24(4):1045–1064
17. Chang J, Gao C, Zheng Y, Hui Y, Niu Y, Song Y, Jin D, Li Y (2021) Sequential recommendation with graph neural networks. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, pp 378–387
18. Wei Y, Wang X, Nie L, He X, Chua T-S (2020) Graph-refined convolutional network for multimedia recommendation with implicit feedback. In: Proceedings of the 28th ACM international conference on multimedia, pp 3541–3549
19. Liu Y, Liu Q, Tian Y, Wang C, Niu Y, Song Y, Li C (2021) Concept-aware denoising graph neural network for micro-video recommendation. In: Proceedings of the 30th ACM international conference on information & knowledge management, pp 1099–1108
20. Kabbur S, Ning X, Karypis G (2013) Fism: factored item similarity models for top-n recommender systems. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, pp 659–667
21. He R, McAuley J (2016) Fusing similarity models with Markov chains for sparse sequential recommendation. In: 2016 IEEE 16th international conference on data mining (ICDM). IEEE, pp 191–200
22. Rendle S, Freudenthaler C, Schmidt-Thieme L (2010) Factorizing personalized markov chains for next-basket recommendation. In: Proceedings of the 19th international conference on World Wide Web, pp 811–820
23. Wang P, Guo J, Lan Y, Xu J, Wan S, Cheng X (2015) Learning hierarchical representation model for nextbasket recommendation. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pp 403–412
24. Hidasi B, Karatzoglou A, Baltrunas L, Tikk D (2016) Session-based recommendations with recurrent neural networks. In: International conference on learning representations (ICLR)
25. Wu S, Tang Y, Zhu Y, Wang L, Xie X, Tan T (2019) Session-based recommendation with graph neural networks. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 346–353
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems, vol 30, pp. 197–206
27. Kang W-C, McAuley J (2018) Self-attentive sequential recommendation. In: 2018 IEEE international conference on data mining (ICDM). IEEE, pp 197–206
28. Zhao W, Wang B, Ye J, Gao Y, Yang M, Chen X (2018) Plastic: prioritize long and short-term information in top-n recommendation using adversarial training. In: Ijcai, pp 3676–3682
29. Gu P, Han Y, Gao W, Xu G, Wu J (2021) Enhancing session-based social recommendation through item graph embedding and contextual friendship modeling. Neurocomputing 419:190–202
30. Hu L, Li C, Shi C, Yang C, Shao C (2020) Graph neural news recommendation with long-term and short-term interest modeling. Inf Process Manag 57(2):102142
31. An M, Wu F, Wu C, Zhang K, Liu Z, Xie X (2019) Neural news recommendation with long-and short-term user representations. In: Proceedings of the 57th annual meeting of the Association for Computational Linguistics, pp 336–345
32. Yu Z, Lian J, Mahmoody A, Liu G, Xie X (2019) Adaptive user modeling with long and short-term preferences for personalized recommendation. In: IJCAI, pp 4213–4219
33. Zheng Y, Gao C, Chang J, Niu Y, Song Y, Jin D, Li Y (2022) Disentangling long and short-term interests for recommendation. In: Proceedings of the ACM Web conference 2022, pp 2256–2267
34. Wu C, Wu F, Qi T, Li C, Huang Y (2022) Is news recommendation a sequential recommendation task? In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp 2382–2386
35. Lathia N, Hailes S, Capra L, Amatriain X (2010) Temporal diversity in recommender systems. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, pp 210–217
36. Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2012) Bpr: Bayesian personalized ranking from implicit feedback. arXiv:1205.2618
37. He X, Liao L, Zhang H, Nie L, Hu X, Chua T-S (2017) Neural collaborative filtering. In: Proceedings of the 26th international conference on World Wide Web, pp 173–182

38. Zhou C, Bai J, Song J, Liu X, Zhao Z, Chen X, Gao J (2018) Atrank: an attention-based user behavior modeling framework for recommendation. In: Thirty-second AAAI conference on artificial intelligence
39. Zhang Y, Dai H, Xu C, Feng J, Wang T, Bian J, Wang B, Liu T-Y (2014) Sequential click prediction for sponsored search with recurrent neural networks. In: Proceedings of the AAAI conference on artificial intelligence, vol 28
40. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
41. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M (2016) Tensorflow: a system for large-scale machine learning. In: 12th $\{USENIX\}$ symposium on operating systems design and implementation ($\{OSDI\}$ 16), pp 265–283
42. Lu Y, Huang Y, Zhang S, Han W, Chen H, Zhao Z, Wu F (2021) Multi-trends enhanced dynamic micro-video recommendation. arXiv:2110.03902
43. Shang Y, Gao C, Chen J, Jin D, Wang M, Li Y (2023) Learning fine-grained user interests for micro-video recommendation. In: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, pp 433–442