MRT: Multi-modal Short- and Long-range Temporal Convolutional Network for Time-sync Comment Video Behavior Prediction

Weihao Zhao[1], Weidong He[1], Hao Wang[1], Haoyang Bi[1], Han Wu[2], Chen Zhu[2,3], Tong Xu[1], Enhong Chen[1] †

[1]University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
[2]Career Science Lab, BOSS Zhipin
[3]School of Management, University of Science and Technology of China
{zhaoweihao, hwd, bhy0521}@mail.ustc.edu.cn;
{ustcwuhan, zc3930155}@gmail.com; {wanghao3, tongxu, cheneh}@ustc.edu.cn

Abstract

As a fresh way to improve the user viewing experience, videos of time-sync comments have attracted a lot of interest. Many efforts have been made to explore the effectiveness of time-sync comments for various applications. However, due to the complexity of interactions among users, videos, and comments, it still remains challenging to understand users' behavior on time-sync comments. Along this line, we study the problem of time-sync comment behavior prediction with considerations of both historical behaviors and multi-modal information of visual frames and textual comments. Specifically, we propose a novel **M**ulti-modal short- and long-**R**ange **T**emporal Convolutional Network model, namely **MRT**. Firstly, we design two amplified Temporal Convolutional Networks with different sizes of receptive fields, to capture both short- and long-range surrounding contexts for each frame and time-sync comments. Then, we design a bottle-neck fusion module to obtain the multi-modal enhanced representation. Furthermore, we take the user preferences into consideration to generate the personalized multi-model semantic representation at each timestamp. Finally, we utilize the binary cross-entropy loss to optimize MRT on the basis of users' historical records. Through comparing with representative baselines, we demonstrate the effectiveness of MRT and qualitatively verify the necessity and utility of short- and long-range contextual and multi-modal information through extensive experiments.

1. Introduction

With the advancement of social media, online videos have become an essential part of human daily lives. Indeed, merely watching videos no longer suffices people's requirements; they would also like to express their opinions and engage in discussions through comments. Recently, a new kind of dynamic comment, named "time-sync comments" (He et al., 2017), has become increasingly popular, especially among young people. As illustrated in Figure 1, users send time-sync comments (e.g. "sneak attack" and "naughty guy") when watching two cats playing, which appear like bullets across the screen simultaneously. In fact, these comments not only enhance the semantic content of the video but also provide a more engaging and interactive experience for users.

As far as we are concerned, several studies have attempted to explore the effectiveness of the time-sync comments for various applications, such as event detection (Li et al., 2016), spoiler detection (Yang et al., 2019b) and comment generation (Wang et al., 2020). Besides, some research works utilized time-sync comments to generate video tags (Lv et al., 2016). In recent years, using time-sync comments for video popularity prediction (He et al., 2016b), video recommendation (Ping, 2018), and video analysis (Pan et al., 2022) has become a hot topic. However, existing research on users' time-sync comment behavior has not been fully explored, limiting the potential to enhance users' interactive experience of watching videos. In fact, accurately predicting users' time-sync comment behavior is essential for enhancing content recommendations, thereby improving users' interactive experience. This enhanced experience can, in turn, encourage users to produce more time-sync comments, enriching the video's quality and boosting overall engagement. Thus, a comprehensive understanding of user time-sync comment behavior is of utmost importance.

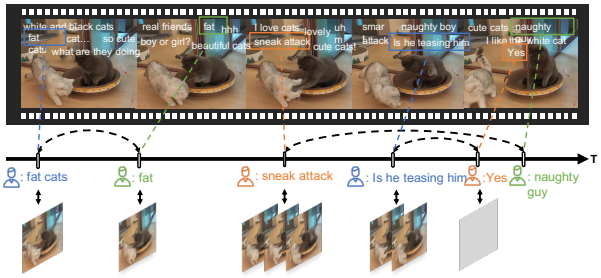


Figure 1: An example of time-sync comment video.

However, there are several technical challenges to this problem. Specifically, 1). *Short- and long-range of semantics*: The semantics represented by the time-sync comment visual frame at each times-

†*CorrespondingAuthor.*

LREC-COLING 2024, pages 11680–11691

tamp are not independent, while instead, related to different ranges of the surrounding contexts, which is intuitive that the contents of the visual frames are continuous and the time-sync comments are also interactive with each other. However, different ranges of contexts can correspond to inconsistent semantic relationships. Here, we take a toy example as shown in Figure. 1. In this video, the time-sync comment "Is he teasing him" is not only relevant with the short-range comment "yes", but also has a correlation with the long-range comment "fat cats", which refers to the word "him". Therefore, it's necessary to distinguish the short-range and long-range surrounding information and extract the different semantics for both visual frames and time-sync comments; 2). *The limitation of the single modality*: The presented contents by the visual frame or time-sync comment at each timestamp are insufficient to infer the complete semantics dependently since it lacks coherence or prior knowledge. Consequently, how to fuse the multi-modal information of visual frames and textual time-sync comments for generating the enhanced representations at each timestamp is a non-trivial problem; 3). *Personalized semantic representation*: There's a common phenomenon on the video platform that users often have different comment behaviors, such as some users would like to interact immediately with current contents, while others tend to summarize previous parts to present the conclusive comments finally. It means that different users focus on different ranges of semantic content, and behavior preferences play an important role in personalized semantic representation for different users.

Along this line, we present a focused study on the problem of time-sync comment behavior prediction with considerations of both historical behavior records and multi-modal information of textual time-sync comments and visual frames. We propose a novel **M**ulti-modal short- and long- **R**ange **T**emporal Convolutional Network model, namely MRT. To be specific, we leverage the amplified Temporal Convolutional Network with different sizes of receptive fields, to capture the short-range and long-range semantics of the surrounding contextual information for each frame and time-sync comment, respectively. Then, considering the mutual correlations between visual frames and textual time-sync comments, we introduce a bottle-neck double-short (long) fusion module to integrate these two ranges of multi-modal information with novel bottle-neck and scale-dot attention mechanisms. Moreover, to capture user preferences, we introduce a Personalized Two-range Multi-modal Information Fusion module to integrate both ranges of information while considering user behavior preference, creating personalized semantic representations at each timestamp. Subsequently, given the historical behavior records and final generated representations at each timestamp in videos, we utilize the binary cross-entropy loss to optimize our proposed MRT model. In addition, we demonstrate the effectiveness of our proposed MRT by comparing it with several representative baselines on a large real-world dataset, and conduct complete ablation studies with several variants of MRT to verify not only the utility of each component but also the necessity of the short- and long-range contextual and multi-modal information.

2. Related Work

2.1. Time-sync video comment

Time-sync comments, offering enhanced real-time user engagement, have garnered increasing attention in research. Early efforts primarily concentrated on annotating videos or video segments. Wu *et al.* (Wu et al., 2014) extracted time-sync video tags by leveraging crowdsourced comments, solving user bias and sparse comments through a temporal and personalized topic model. Yang *et al.* (Yang et al., 2017b) proposed SW-IDF, an unsupervised video tag extraction algorithm, leveraging semantic association graphs derived from time-sync comments to differentiate meaningful comments from noise. Lv *et al.* (Lv et al., 2016) introduced a framework that assigned temporal labels to highlighted video shots by representing time-sync comments and recognizing video highlights via semantic vectors in a supervised manner. Lately, efforts on additional applications utilizing time-sync comments have emerged. Ping *et al.* (Ping and Chen, 2017) focused on video highlight detection using concept-mapped lexical chains for lag calibration, modeling video highlights based on comment intensity and emotion-concept concentration. Li *et al.* (Li et al., 2016)proposed a model for event detection using Time-Sync comments, extracting features from comments, and analyzing user behavior relevance. Yang *et al.* (Yang et al., 2019b) designed the Similarity-Based Network with Interactive Variance Attention (SBN-IVA) to classify time-sync comments as spoilers or not. Xu *et al.* (Xu and Zhang, 2017) generated temporal descriptions of videos using crowdsourced time-sync comments, addressing the challenge of informal and noisy comments by selecting representative ones based on a temporal summarization model. Bai *et al.* (Bai et al., 2021) addressed the task of aligning time-sync video comments to narrative video storylines, utilizing variational auto-encoders to map comments and storylines into latent spaces and applying dynamic programming for global optimal outputs. Hu *et al.* (Hu et al., 2022) proposed a method for classifying time-sync comment videos by combining the BERT model with a machine-learning classifier to

analyze TSCs and titles. (Pan et al., 2022) focused on temporal information capture in affective video content analysis by leveraging time-synchronized comments as auxiliary supervision. Further usages of the time-sync comment were proposed, such as video recommendation (Ping, 2018; Zhao et al., 2023), comment generation (Ma et al., 2019; Wang et al., 2020) and video emotion analysis (Cao et al., 2022).

2.2. User behavior analysis in videos

Understanding users' behaviors when watching online videos is crucial to the design of online video platforms. Early work focused on user behaviors, content access patterns, and their implications on the design of online video systems (Yu et al., 2006; Mongy et al., 2005). Along this line, Qiu *et al.* (Qiu and Cui, 2010) extended this research to the micro level of the individual and categorized the viewers' behaviors of watching online videos into seven patterns. Recently, researchers have concentrated on predicting the popularity of videos. For example, Chen *et al.* (Chen et al., 2018) proposed a fine-grained video attractiveness prediction using multi-modal deep learning on a large real-world dataset, and Huang *et al.* (Huang et al., 2018) made a focused study on user behavior analysis and video popularity prediction. With the advancement of recommendation system technology, an increasing number of works utilize user behavioral characteristics for information filtering (Wang et al., 2019; Wu et al., 2019b; Yang et al., 2019a; Wang et al., 2021), leading to a diverse range of applications in the video domain. Yang *et al.* (Yang et al., 2017a) explored modeling user preferences across multiple video websites, proposing a Multi-site Probabilistic Factorization (MPF) model to capture both cross-site and site-specific interests based on viewing records from a large ISP. Jiang *et al.* (Jiang et al., 2020) proposed an end-to-end Multi-scale Time-aware user Interest Modeling Network (MTIN) for micro-video recommendation. Wu *et al.* (Wu et al., 2019c) proposed to project both users and items into a latent collaborative space and a visual space for the personal key frame recommendation. Chen *et al.* (Chen et al., 2021) took the diversity of users' interests into account and proposed a user preference reasoning method to predict frame-level preferences. However, as far as we are concerned, few works considered studying the behavior prediction problem of sending time-sync comments from the perspective of multi-modal modeling.

3. Problem Definition

We denote $U = \{u_1, ..., u_M\}$ as a set of users and $H = \{h_1, ..., h_N\}$ as a series of videos on the service platform, where M and N indicate the number of users and videos, respectively. Specifically, each video h_j consists of multiple frames and multiple time-sync comments, $h_j = (A_j, C_j)$, where A_j, C_j stands for the frames set and the time-sync comments set of j-th video. $A_j = \{a_j^{(t)}\}_{t=1}^{T}$, where $a_j^{(t)}$ denotes the frames at different time t in A_j, and every frame $a_j^{(t)}$ is especially associated with a list of time-sync comments $\{c_{l|j}^{(t)}\}_{l=1}^{L}$ of length L at the corresponding time t. Since the length of each video and time-sync comments per frame are not exactly the same, the number of video timestamps T and time-sync comments L will vary in different videos. Besides, the service platform also collects lots of user history behaviors $R_{ij}^{(t)}$, where the element of records is a binary value. $R_{ij}^{(t)} = 1$ indicates that user u_i has previously commented on the t-th frame $a_j^{(t)}$ of video h_j, and vice versa. To this end, the problem of this paper can be formulated as follows:

Definition 1 *Given the users U, videos H containing frames A and time-sync comments C, and user history behaviors $R_{ij}^{(t)}$, we study a novel time-sync comment behavior prediction problem in this paper, which aims to predict the user behaviors $\hat{R}_{ij}^{(t)}$ on non-interactive videos.*

4. Technical Details

In this section, we will introduce the architecture of our proposed model (MRT) shown in Figure 2.

4.1. Initial Multi-modal Feature Extraction

As depicted in Figure 2, the input contains the data from two modalities: visual frames and textual comments. Following previous works (Badamdorj et al., 2021; Chen et al., 2021), we extract initial representations of visual frames and textual comments with the pre-trained models, ResNet (He et al., 2016a) and BERT (Devlin et al., 2018), denoted as $A = \{a_t\}_{t=1}^{T}$ and $C = \{c_t^L\}_{t=1}^{T}$ respectively. Considering the variable length L of time-sync comments $\{c_{l|j}^{(t)}\}_{l=1}^{L}$ at different timestamps $a_j^{(t)}$, we first pad the lengths of all time-sync comments to the same size. Then, we convert the aligned time-sync comments tensor $C \in \mathbb{R}^{T\times L\times d}$ to matrix $C \in \mathbb{R}^{T\times(L\times d)}$, which will speed up the mini-batch training procedure. To maintain the uniqueness of each frame while facing the feature content lost, we concatenate the frame ID embedding matrix with the visual frames matrix A and the time-sync comments matrix C, respectively. Especially, a linear transformation layer is leveraged to unify the feature dimension. In this way, the visual frames and textual time-sync comments will be transformed into representation matrix $A \in \mathbb{R}^{T\times d}$ and $C \in \mathbb{R}^{T\times d}$,

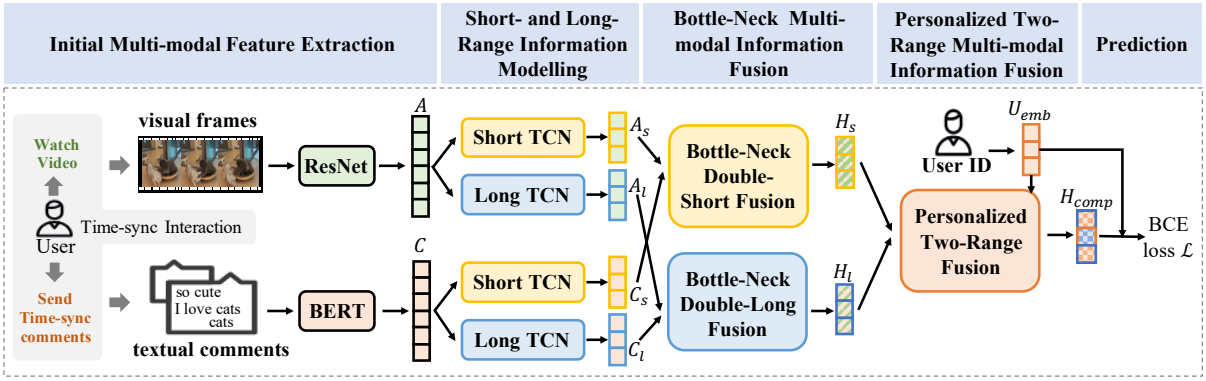


Figure 2: The framework of **M**ulti-modal short and long **R**ange **T**emporal Convolutional Network (**MRT**).

where the notation d indicates the corresponding embedding dimensions.

4.2. Short- and Long-range Information Modeling

Videos with time-sync comments are a form of dynamic streaming data, leading to the following two phenomena. First, the video content is continuous, with a complete segment being collectively expressed by a range of frames. Therefore, the content of each frame is interrelated with the surrounding frames. Second, the semantics of each frame rely on both short- and long-range information. Short-range is immediate and tied to timestamps, whereas long-range offers a general and contextual background. It's crucial to integrate these different ranges for complete understanding of both time-sync comments and visual frames.

To address these issues, we design two modified Temporal Convolutional Networks (TCN) (Bai et al., 2018; Lea et al., 2017), named Short TCN and Long TCN, which are tailored to capture the specific ranges of information for both single-modal time-sync comments and frames. Along this line, we employ different sizes of dilated convolution factors to obtain different receptive fields efficiently and define the dilated convolution operation on each element e of the input sequence $x \in \mathbb{R}^n$ with a convolution filter $f : \{0, ..., k-1\} \rightarrow \mathbb{R}$ as:

$$\text{Conv}(e) = (x *_b f)(e) = \sum_{i=0}^{k-1} f(i) \cdot x_{e-b \cdot i}\ , \quad (1)$$

where k denotes the kernel size, subscript $e - b \cdot i$ accounts for the direction of the past, and b denotes the factor of dilation. The architecture of Short (Long) TCN contains a series of blocks, each block consists of one dilated convolution layer with a ReLU activation layer, a dropout layer, and a subsequent 1×1 convolution layer with the same ReLU and dropout layer. For the short-range information, as illustrated in Figure 3, we utilize a small kernel size with a dilation factor b of 1 for each dilation layer to emphasize instance-level semantics. For the long-range information, as shown on the right side of Figure 3, we take a larger kernel size at the first dilation convolution layer, and follow-up reduces the kernel size but increases the dilation factor b exponentially with the depth of the network. Then, we stack numbers of blocks with residual connections to formalize the final short-range and long-range TCN_s and TCN_l, obtaining representations of different range contexts by:

$$C_s = \text{TCN}_s^C(C), \quad (2)$$

$$C_l = \text{TCN}_l^C(C). \quad (3)$$

In this way, we generate two different representations C_s and C_l for the time-sync comments. Note that, we finely design the kernel size and the dilation factor to make sure the largest receptive fields of the short-range TCN are smaller than the minimum receptive field of the long-range TCN. Here, C_s mainly focuses on specific and detailed supplements to the current timestamp, while C_l considers the long-range that represents more general or high-level semantic information, since TCN_l has larger receptive fields than TCN_s.

Furthermore, we adopt a similar architecture to obtain representations of short-range and long-range contextual information for the visual frame:

$$A_s = \text{TCN}_s^A(A), \quad (4)$$

$$A_l = \text{TCN}_l^A(A). \quad (5)$$

Finally, we could derive the short-range and long-range representations A_s, C_s, A_l, and C_l for both visual frames and time-sync comments for a video from Eq. 2 $\sim$ Eq. 5.

4.3. Bottle-Neck Multi-modal Information Fusion

In time-sync comment videos, users often comment on semantic contents that aligns with their

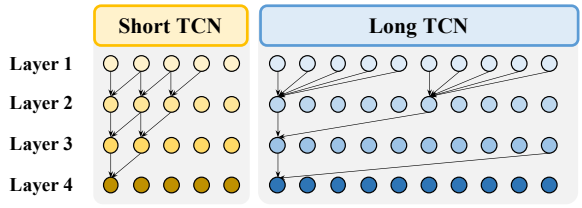


Figure 3: The Modified Short and Long TCN.

interests. However, these semantic contents, such as plot and story logic, can be intricate and abstract, presenting a significant challenge for representation using single-modal information. As time-sync comments are typically brief and concise, visual information could provide additional contexts for the mentioned contents, while the detailed information within time-sync comments could enrich the semantics of frames that are difficult to extract directly from visual observations. Therefore, after deriving short- and long-range representations for both visual frames and time-sync comments, we intend to integrate them to generate comprehensive representations of contents at each timestamp.

To this end, we propose a Bottle-Neck Double-Short (Long) Fusion Module to integrate multi-modal information, inspired by the bottle-neck mechanism (Nagrani et al., 2021), we compress two single-modal information flows through a tight bottle-neck, which forces the model to condense information from each modality. Specifically, for the aforementioned short-range representations of visual frames and time-sync comments, we initially combine them together and project into a lower dimension co-representation space, which is defined as the following *CO-Matrix* M_s:

$$M_s = [A_s \oplus C_s]W_M^s, \tag{6}$$

where $\oplus$ denotes the concatenation of two vectors, $W_M^s \in \mathbb{R}^{2d \times d_{\text{bottle}}}$ refers to trainable parameters, d is the feature dimension size, and d_{bottle} is the compress dimension with the bottle-neck mechanism. Note that, we set a smaller dimension size of d_{bottle} to compress the single-modal information flow for extracting condensed information from each modality. Next, we introduce the scale-dot attention mechanism (Vaswani et al., 2017) to design an elaborated fusion, which incorporates the compressed information to generate more comprehensive representations, illustrated as:

$$K_{M_s}^A = M_s W_K^A, K_{M_s}^C = M_s W_K^C, \tag{7}$$

$$V_{M_s}^A = M_s W_V^A, V_{M_s}^C = M_s W_V^A. \tag{8}$$

Here, the notations $W_K^A, W_K^C, W_V^A, W_V^A \in \mathbb{R}^{d \times d_{\text{bottle}}}$ are trainable parameters. Different from pairwise or cross attention mechanism using the concatenated feature M_s as the *Query* vector, we take the compressed information to obtain the *Key* and *Value* vectors in the scale-dot attention mechanism, to calculate the attention scores between single-modal information A_s, C_s and the multi-modal information M_s, respectively. Then, we get the multi-modal information enhanced representation A_s^{attn} and C_s^{attn} by the weight accumulation operation as follows:

$$\begin{aligned} Q_s^A &= A_s W_Q^A, Q_s^C = C_s W_Q^C, \\ A_s^{\text{attn}} &= \text{softmax}(\frac{Q_s^A (K_{M_s}^A)^{\text{T}}}{\sqrt{d}}) V_{M_s}^A, \\ C_s^{\text{attn}} &= \text{softmax}(\frac{Q_s^C (K_{M_s}^C)^{\text{T}}}{\sqrt{d}}) V_{M_s}^C. \end{aligned} \tag{9}$$

The $W_Q^A, W_Q^C \in \mathbb{R}^{d \times d}$ are the linear projection matrix which convert the input vector A_s and C_S into the *Query Matrix* Q_s^A and Q_s^C. After getting the enhanced representation A_s^{attn} and C_s^{attn}, we feed them into a feed-forward network, and define the mathematical calculation as follows:

$$A_s^{'} = \text{FFN}(A_s^{\text{attn}}), \tag{10}$$

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2, \tag{11}$$

where $W_1, W_2 \in \mathbb{R}^{d \times d}$ and $b_1, b_2 \in \mathbb{R}^{1 \times d}$ are trainable parameters in the feed-forward network. With this designed combination of the bottle-neck and scale-dot mechanism, we promote the condensed vector to retain the most meaningful information and abort the irrelevant or redundant information in a single modality. Subsequently, we bolster the scale-dot mechanism with the condensed vector to produce improved representations of short-range information. After stacking several blocks, we concatenate the enhanced features A_s^e and C_s^e to obtain the final fusion representation H_s of both short-range visual frames and time-sync comments as:

$$H_s = \text{ReLU}([A_s^e \oplus C_s^e]W_H^s), \tag{12}$$

where ReLU is a non-linear activate function, and $W_H^s \in \mathbb{R}^{2d \times d}$ is the trainable parameters.

Similarly, for long-range information, we adopt the same architecture as defined in the short-range information fusion to generate the comprehensive representation H_l at each timestamp.

4.4. Personalized Two-range Multi-modal Information Fusion

After obtaining the comprehensive representations of multi-modal H_s and H_l, the next step is to combine them together to represent the content semantics at each timestamp. However, a common phenomenon on the video platform is that different users have different behavior preferences. Some users would like to comment on a specific context based on a certain frame, while others tend to summarize multiple previous timestamp content to

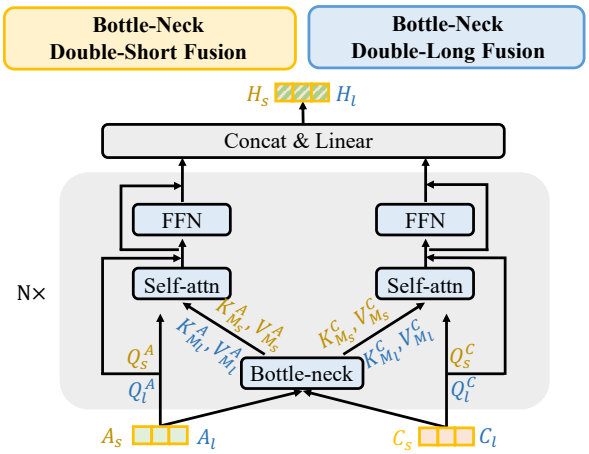


Figure 4: Bottle-Neck Double-Short (Long) Fusion.

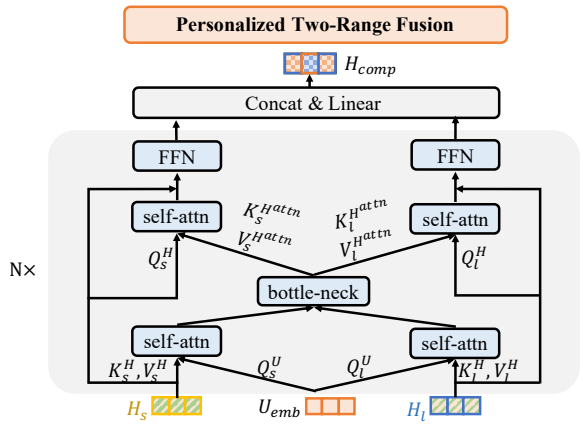


Figure 5: The Personalized Two-Range Fusion.

express conclusive opinions. Therefore, it is necessary to take the user behavior preferences into consideration to generate personalized semantic representations at each timestamp.

To this end, we propose a Personalized Two-range Multi-modal Information Fusion module, which combines short-range and long-range information while also considering the user behavior preference, as illustrated in Figure 5. Specifically, we convert the user embeddings U_{emb} into different *Query* matrices Q_s^U and Q_l^U and formulate the obtained representations H_s and H_l to the corresponding *Key* matrices K_s^H and K_l^H in the scale-dot mechanism, to distinguish the user behavior preferences with respect to the short- and long-range information.

$$Q_s^U = U_{\text{emb}} W_Q^s, Q_l^U = U_{\text{emb}} W_Q^l, \quad (13)$$

$$K_s^H = H_s W_K^s, K_l^H = H_l W_K^l, \quad (14)$$

$$V_s^H = H_s W_V^s, V_l^H = H_l W_V^l, \quad (15)$$

$$H_s^{\text{attn}} = \text{softmax}(\frac{Q_s^U (K_s^H)^{\text{T}}}{\sqrt{d}}) V_s^H, \quad (16)$$

$$H_l^{\text{attn}} = \text{softmax}(\frac{Q_l^U (K_l^H)^{\text{T}}}{\sqrt{d}}) V_l^H, \quad (17)$$

where $W_Q^s, W_Q^l, W_k^s, W_K^l, W_V^s, W_V^l \in \mathbb{R}^{d \times d}$ are the trainable parameters. With the calculated attentions and weight accumulation defined in Eq. 16 and Eq. 17, we can obtain the personalized enhanced representations H_s^{attn} and H_l^{attn} of short-range and long-range at each timestamp. Similar to the previous bottle-neck module, we also combine the H_s^{attn} and H_l^{attn} as the input and compress them into a compact representation space to filter out irrelevant parts between the watching users, short-range and long-range information:

$$M_H = [H_s^{\text{attn}} \oplus H_l^{\text{attn}}] W_M^H, \quad (18)$$

where the trainable $W_M^H \in \mathbb{R}^{2d \times d_{\text{bottle}}}$ forces to generate the compress vector M_H. Then, the compact fusion representation is transformed into *Key*,*Value* matrix and associated with *Query* matrix derived from the H_s^{attn} and H_l^{attn}, to obtain the enhanced representations for both personalized short-range and long-range context information respectively, which has been defined in detail by previous Eq. 7∼ Eq. 11. Therefore, we omit the same formulation here for brevity and directly denote the two ranges of calculated personalized enhanced representations as H_s^e and H_l^e, which are naturally combined together and following with the trainable parameters $W_H^{\text{comp}} \in \mathbb{R}^{d \times d}$ and activation function ReLU to obtain the final fusion representation H_{comp} at each timestamp in a video as:

$$H_{\text{comp}} = \text{ReLU}([H_s^e \oplus H_l^e] W_H^{\text{comp}}). \quad (19)$$

4.5. Objective Function

Once we obtain the final representations of each frame in a video H_{comp}, we multiply the user embedding vector U_{emb} with H_{comp} and leverage the Sigmoid function to predict the plausibility score on the i-th frame of the video as $s_i = \sigma(U_{\text{emb}} * H_{\text{comp}})$, where s_i indicates the probability of the user's time-sync comment behavior on the target frame. The final objective function is formulated as:

$$\mathcal{L} = \sum_{i=1}^{T} -[r_i \log(s_i) + (1 - r_i) \log(1 - s_i)], \quad (20)$$

where $r_i \in \{0, 1\}$ denotes the ground truth of users' time-sync comment records.

5. Experiments

5.1. Experimental Setting

Dataset. The data provided by Lv *et al.* (Lv et al., 2019) is constructed from a real-world time-sync comment-enabled video-sharing platform. To fit it

Table 1: The performance of different methods on time-sync comment behavior prediction.

Methods	NDCG@1	NDCG@5	NDCG@10	Recall@1	Recall@5	Recall@10	mAP
Random	0.0230	0.0538	0.0823	0.0172	0.0851	0.1696	0.0819
MostPopular	0.0274	0.0652	0.0987	0.0219	0.1016	0.2006	0.0926
PMF	0.0336	0.0720	0.1049	0.0251	0.1109	0.2080	0.0976
BPR	0.0350	0.0755	0.1095	0.0261	0.1170	0.2175	0.1006
JIFR	0.0298	0.0669	0.0988	0.0219	0.1047	0.1989	0.0930
JIFR+T	0.0372	0.0846	0.1289	0.0272	0.1328	0.2640	0.1155
ITF-HEA	0.0402	0.0890	0.1329	0.0308	0.1372	0.2683	0.1204
MRT(Ours)	**0.0827**	**0.1660**	**0.2217**	**0.0631**	**0.2487**	**0.4126**	**0.1823**

Table 2: The ablation models for comparison.

Target Module	Acronym	Replacement
Personalized Two-range Multi-modal Information Fusion	**NoRaFu**	simple concatenate operation
	RaSeFu	self attention architecture
	RaUsFu	user attention architecture
Bottle-Neck Multi-modal Information Fusion	**NoMuFu**	simple concatenate operation
	MuSeFu	self attention architecture
Short-and Long-Range Information Modelling	**NoShort**	remove short-range
	NoLong	remove long-range
Different modalities	**NoImage**	remove visual modality
	NoText	remove textual modality

with the current task, we preprocessed the original dataset as follows. First, we deleted videos that were less than 10 minutes. After that, we used an automatic tool [1] to clip the video into smaller segments. Without ambiguity, we still refer to these segments as videos. Due to the uneven distribution of the time-sync comments, we filtered out the videos which have less than one comment per frame on average. To obtain enough information to model the users' interactive behaviors, we selected users who have interacted with more than 10 videos. Finally, the preprocessed dataset contains 9,668 users and 10,629 videos, 694,724 frames, and 1,377,780 time-sync comments. Note that time-sync comments that are not produced by the target user are also included since they are helpful in comprehending the semantics of video frames. For the target users, we have 236,395 interaction records, and we randomly split 70% of the data for training, 10% for validation, and 20% for testing.

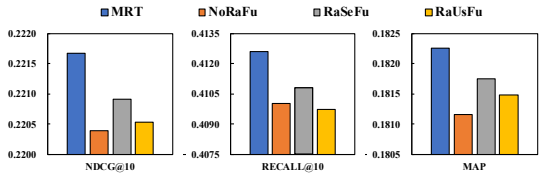
Baselines. We adopt two types of baselines for performance comparison. The first category is based on collaborative filtering, including Random, Mostpopular, PMF (Mnih and Salakhutdinov, 2007), and BPR (Rendle et al., 2009). In particular, Random gives randomly sets the probability, while Mostpopular sets the probability positively related to the historical popularity of the frame. PMF is frequently adopted for rating-based prediction tasks, and the BPR is a ranking-based approach for user implicit feedback modeling. The second category leverages side information for the prediction, including JIFR, its variant JIFR+T (Wu et al., 2019c), and ITF-HEA (Yang et al., 2019a). Specifically, JIFR is designed for personalized multimedia item and key frame recommendation, and the JIFR+T is a variant of JIFR specifically implemented for our experiment, which represents each frame with the help of the corresponding time-sync comments. ITF-HEA is a state-of-the-art model that jointly incorporates the context-dependent and time-sensitive properties of both time-sync comments and visual frames.

Evaluation Protocols and Parameter Setting. We apply three metrics for evaluation: NDCG@K, Recall@K, and mAP, and take the K values of $\{1, 5, 10\}$ for a comprehensive evaluation. We set the size of the hidden vector d as 128, and the size of the bottle hidden vector d_{bottle} as 32. The number of blocks for each fusion module is 2 and the number of blocks for short and long TCN is 4. The kernel size of the short TCN is 2 and the dilation factor is 1 at each block. As for long TCN, we set the kernel size as 5 and the dilation factor as 1 in the first block. After that, we set the kernel size as 2 and set the dilation factor as $5 \cdot 2^{i-1}$ in the i-th blocks. During training, we use Adam (Kingma and Ba, 2014) as our optimizer and set the learning rate as 0.00005 and the mini-batch size as 32.

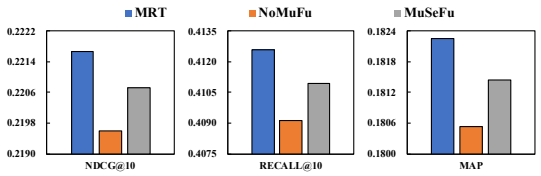
5.2. Overall Performance

We present the overall performance in Table 1. We can obviously find that MRT outperforms all the baselines by a large margin on all metrics. Among the baselines, the superiority of MostPopular over Random provides some degree of corroboration for the "Herding effect" of the time-sync comment behavior, as mentioned in (Yang et al., 2019a). This suggests that using a set of synchronized comments within a given scope enhances the semantic representation of the current frame compared to isolated comments. PMF is superior to Random and MostPopular, as PMF leverages collaborative filtering to extract useful information from user-frame interaction records. BPR achieves expected results by directly optimizing pairwise ranking loss, to enables better capture of the item ranking relationships. However, MRT performs significantly better than the aforementioned methods, attributed to the

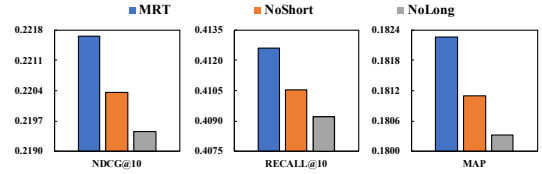
[1] https://screenpy-docs.readthedocs.io/en/latest/



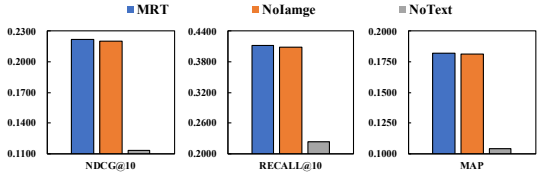
(a) The comparison between Personalized Two-Range Fusion and its variants.



(b) The comparison between Bottle-Neck Multi-modal Information Fusion and its variants.



(c) The performance of Short- and Long-Range Information Modelling and its variants.



(d) The performance of MRT comparing between different modality inputs.

Figure 6: Evaluation of MRT compared with different ablation models

complex and sparse nature of user-barrage frame interactions. In summary, MRT provides a new modeling perspective based on semantic content rather than only interaction records to better capture the relationships between users and frames.

However, JIFR, which also focuses on contents, performs worse than PMF and BPR. This is due to that users' commenting behavior does not simply describe the current video frame, but interacts with the higher-level semantic information conveyed by frames. Such high-level semantics are difficult to obtain solely from the visual modality. In contrast to JIFR, JIFR-T has achieved a significant improvement and outperformed PMF and BPR models. On one hand, JIFR-T utilizes textual information to construct the semantic representation, confirming the effectiveness of the textual modality. On the other hand, the textual information is inherently more abstract and is closer to the high-level semantic information carried by video frames, hence resulting in better performance of JIFR-T. However, MRT outperforms JIFR and JIFR-T, which could be due to their dependence on extracting semantic information solely from individual frames and comments while falling short of capturing the complex high-level information conveyed by the frame sequence.

ITF-HEA achieves better performance than other baselines, likely because of its consideration of the herd effect in user comment behavior. However, our MRT model outperforms ITF-HEA, which is because ITF-HEA simply concatenates two modalities, neglecting the complex interrelationships between them, while our MRT models the fusion of text and visual modalities thoroughly, enabling the capture of more comprehensive information.

5.3. Ablation Study and Model Analysis

To verify the effectiveness of each component in MRT, we compare it with its variants listed as Table 2. All the experiments are conducted 5 times and the results are the average value. To demonstrate the effectiveness of each module, we also take **the Welch's unequal variances t-test** for all results. The results show a strong statistical significance difference between our model and all variants with a significance level of 0.05 ($p<0.05$).

Figure 6a depicts the comparison between the range fusion module and its variants NoRaFu, RaSeFu, and RaUsFu. We can find MRT outperforms all three variants, indicating that the Personalized Two-Range Multi-modal Information Fusion module can better integrate the short- and long-range information. Though RaSeFu outperforms RaUsFu, it is not as effective as MRT, which confirms the effectiveness of the bottle-neck mechanism in fusing information from different ranges.

Figure 6b shows the comparison between the Bottle-Neck Multi-modal Information Fusion module and its variants, i.e., NoMuFu, MuSeFu. The performance drops a lot when we replace this fusion module with concatenation or self-attention, this demonstrates the effectiveness of our module in integrating the two modalities of information. Moreover, the progressive performance improvement of the NoMuFu, MuSeFu, and MRT models serves as further evidence of our motivation that the relationship between the text and visual modalities is highly intricate and requires a sophisticated model design to capture the underlying inter-modal relationships.

Figure 6c shows the performance of MRT compared with only utilizing short TCN and long TCN. By observing the significant performance decrease when utilizing the single-range TCN, we can verify the motivation of modeling the semantics of a video frame together with its surrounding frames from different ranges. Besides, the performance improved when using long-range information compared to short-range information. This is because

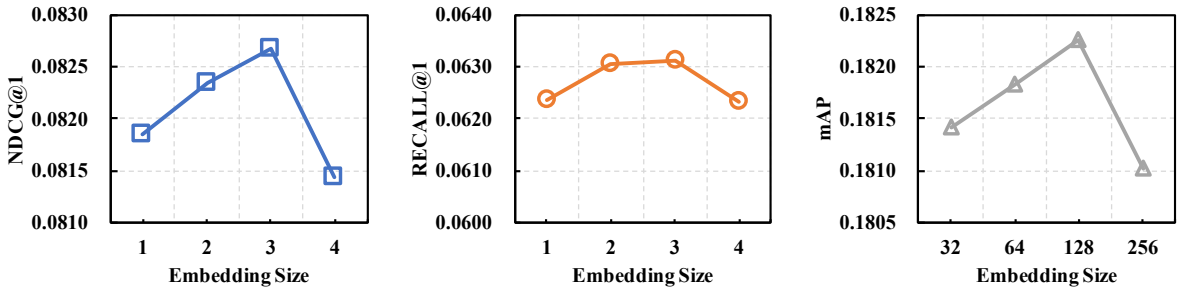


Figure 7: Results of MRT with variant dimensions.

long-range one captures broader trends in user commenting behavior, confirming that the user comments interact with higher-level semantic information in the video, going beyond simple descriptions.

Figure 6d shows the comparison between MRT and different modality inputs NoImage and NoText. The results confirm the effectiveness of incorporating multi-modal information compared to relying solely on single modal. Furthermore, NoImage outperforms NoText significantly, indicating that textual information is more effective, which is consistent with the results obtained by JIFR and JIFR-T. Firstly, users tend to comment on the high-level semantics rather than on the low-level visual information. This is exemplified by the two cats in Figure 1, where users tend to comment on their *state* (e.g. *fat, cute*) or their *behavior* (e.g. *playing with each other*) rather than just describing their appearance. Extracting such high-level semantics (e.g. *state, behavior*) from visual features is quite difficult. Secondly, textual features can be naturally associated with high-level semantic information due to their abstractness compared to visual features. Finally, user comment behavior is more closely related to the text features since users tend to communicate with others using time-sync comments.

5.4. Parameter Sensitivity

We conduct an additional study to investigate the sensitivity of our modal to the embedding size. Specifically, we compare the performance of our model using different embedding dimensions in the set {32, 64, 128, 256}, and the results are presented in Figure. 7. Our observations reveal that larger embedding dimensions at the beginning contribute to better performance on all metrics, as they can retain more information for training. However, when the dimension exceeds around 128, the performances start to decline, since larger embedding sizes may introduce more noise that could reduce the model's performance.

6. Conclusion

Modeling user time-sync comment behavior is of great significance for inferring user preference for video content. We make a focused study on the user behavior of sending time-sync comments and proposed a novel Multi-modal short- and long-Range Temporal Convolutional Network model (MRT) to solve the prediction problem of time-sync comment behaviors. First, two temporal convolutional networks with different sizes of receptive fields are introduced to capture both short- and long-range contextual. Then, the multi-modal information is integrated through a bottle-neck attention module. After that, the user behavior preferences are utilized to obtain the personalized semantic representation at each timestamp. We demonstrate the effectiveness of MRT on a large real-world dataset and verified the necessity of short- and long-range contextual and multi-modal information. We hope this work will lead to more future studies.

7. Acknowledgements

This work was supported in part by the grants from National Natural Science Foundation of China (No.U23A20319, 62222213, U22B2059, 62072423, 62202443), and the Anhui Provincial Science and Technology Major Project (No. 2023z020006).

8. References

Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. 2021. Joint visual and audio learning for video highlight detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8127–8137.

Qingchun Bai, Yuanbin Wu, Jie Zhou, and Liang He. 2021. Aligned variational autoencoder for matching danmaku and video storylines. *Neurocomputing*, 454:228–237.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Reddy Mounika Bommisetty, Om Prakash, and Ashish Khare. 2020. Keyframe extraction using pearson correlation coefficient and color moments. *Multimedia Systems*, 26(3):267–299.

Wei Cao, Kun Zhang, Han Wu, Tong Xu, Enhong Chen, Guangyi Lv, and Ming He. 2022. Video emotion analysis enhanced by recognizing emotion in video comments. *International Journal of Data Science and Analytics*, 14(2):175–189.

Runnan Chen, Penghao Zhou, Wenzhe Wang, Nenglun Chen, Pai Peng, Xing Sun, and Wenping Wang. 2021. Pr-net: Preference reasoning for personalized video highlight detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989.

Xinpeng Chen, Jingyuan Chen, Lin Ma, Jian Yao, Wei Liu, Jiebo Luo, and Tong Zhang. 2018. Fine-grained video attractiveness prediction using multimodal deep learning on a large real-world dataset. In *Companion Proceedings of the The Web Conference 2018*, pages 671–678.

Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin. 2017a. Personalized key frame recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 315–324.

Yue Chen, Qin Gao, and Pei-Luen Patrick Rau. 2017b. Watching a movie alone yet together: understanding reasons for watching danmaku videos. *International Journal of Human–Computer Interaction*, 33(9):731–743.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gang Fu, Qing Zhang, Qifeng Lin, Lei Zhu, and Chunxia Xiao. 2020. Learning to detect specular highlights from real-world images. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1873–1881.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings.

Eunyoung Han and Sang-Woo Lee. 2014. Motivations for the complementary use of text-based media during linear tv viewing: An exploratory study. *Computers in Human Behavior*, 32:235–243.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ming He, Yong Ge, Enhong Chen, Qi Liu, and Xuesong Wang. 2017. Exploring the emerging type of comment for online videos: Danmu. *ACM Transactions on the Web (TWEB)*, 12(1):1–33.

Ming He, Yong Ge, Le Wu, Enhong Chen, and Chang Tan. 2016b. Predicting the popularity of danmu-enabled videos: A multi-factor view. In *International Conference on Database Systems for Advanced Applications*, pages 351–366. Springer.

Zeyu Hu, Jintao Cui, Wei-Hua Wang, Feng Lu, and Binhui Wang. 2022. Video content classification using time-sync comments and titles. In *2022 7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pages 252–258. IEEE.

Lei Huang, Bowen Ding, Aining Wang, Yuedong Xu, Yipeng Zhou, and Xiang Li. 2018. User behavior analysis and video popularity prediction on a large-scale vod system. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(3s):1–24.

Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3487–3495.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165.

Jiangfeng Li, Zhenyu Liao, Chenxi Zhang, and Jing Wang. 2016. Event detection on online videos using crowdsourced time-sync comment. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pages 52–57. IEEE.

Zhenyu Liao, Yikun Xian, Xiao Yang, Qinpei Zhao, Chenxi Zhang, and Jiangfeng Li. 2018. Tscset: A crowdsourced time-sync comment dataset for exploration of user experience improvement. In *23rd International Conference on Intelligent User Interfaces*, pages 641–652.

Guangyi Lv, Tong Xu, Enhong Chen, Qi Liu, and Yi Zheng. 2016. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Guangyi Lv, Kun Zhang, Le Wu, Enhong Chen, Tong Xu, Qi Liu, and Weidong He. 2019. Understanding the users and videos by mining a novel danmu dataset. *IEEE Transactions on Big Data*.

Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2019. Livebot: Generating live video comments based on visual and textual contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6810–6817.

Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48):12714–12719.

Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20.

Sylvain Mongy, Fatma Bouali, and Chabane Djeraba. 2005. Analyzing user's behavior on a video database. In *Proceedings of the 6th international workshop on Multimedia data mining: mining integrated media and complex data*, pages 95–100.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34.

Jicai Pan, Shangfei Wang, and Lin Fang. 2022. Representation learning through multimodal attention and time-sync comments for affective video content analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 42–50.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Qing Ping. 2018. Video recommendation using crowdsourced time-sync comments. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 568–572.

Qing Ping and Chaomei Chen. 2017. Video highlights detection and summarization with lag-calibration based on concept-emotion mapping of crowd-sourced time-sync comments. *arXiv preprint arXiv:1708.02210*.

Fan Qiu and Yi Cui. 2010. An analysis of user behavior in online video streaming. In *Proceedings of the international workshop on Very-large-scale multimedia corpus, mining and retrieval*, pages 49–54.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461.

Mrigank Rochan, Mahesh Kumar Krishna Reddy, Linwei Ye, and Yang Wang. 2020. Adaptive video highlight detection by learning from user history. In *European Conference on Computer Vision*, pages 261–278. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hao Wang, Defu Lian, Hanghang Tong, Qi Liu, Zhenya Huang, and Enhong Chen. 2021. Hypersorec: Exploiting hyperbolic user and item representations with multiple aspects for social-aware recommendation. *ACM Transactions on Information Systems (TOIS)*, 40(2):1–28.

Hao Wang, Tong Xu, Qi Liu, Defu Lian, Enhong Chen, Dongfang Du, Han Wu, and Wen Su. 2019. Mcne: An end-to-end framework for learning multiple conditional network representations of social network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1064–1072.

Weiying Wang, Jieting Chen, and Qin Jin. 2020. Videoic: A video interactive comments dataset and multimodal multitask learning for comments generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2599–2607.

Bin Wu, Erheng Zhong, Ben Tan, Andrew Horner, and Qiang Yang. 2014. Crowdsourced time-sync video tagging using temporal and personalized

topic modeling. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 721–730.

Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019a. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293.

Le Wu, Lei Chen, Richang Hong, Yanjie Fu, Xing Xie, and Meng Wang. 2019b. A hierarchical attention model for social contextual image recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 32(10):1854–1867.

Le Wu, Lei Chen, Yonghui Yang, Richang Hong, Yong Ge, Xing Xie, and Meng Wang. 2019c. Personalized multimedia item and key frame recommendation. *arXiv preprint arXiv:1906.00246*.

Linli Xu and Chao Zhang. 2017. Bridging video content and comments: Synchronized video description with temporal summarization of crowdsourced time-sync comments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Chunfeng Yang, Huan Yan, Donghan Yu, Yong Li, and Dah Ming Chiu. 2017a. Multi-site user behavior modeling and its application in video recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 175–184.

Wenmian Yang, Wenyuan Gao, Xiaojie Zhou, Weijia Jia, Shaohua Zhang, and Yutao Luo. 2019a. Herding effect based attention for personalized time-sync video recommendation. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 454–459. IEEE.

Wenmian Yang, Weijia Jia, Wenyuan Gao, Xiaojie Zhou, and Yutao Luo. 2019b. Interactive variance attention based online spoiler detection for time-sync comments. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1241–1250.

Wenmian Yang, Na Ruan, Wenyuan Gao, Kun Wang, Wensheng Ran, and Weijia Jia. 2017b. Crowdsourced time-sync video tagging using semantic association graph. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 547–552. IEEE.

Hongliang Yu, Dongdong Zheng, Ben Y Zhao, and Weimin Zheng. 2006. Understanding user behavior in large-scale video-on-demand systems. *ACM SIGOPS Operating Systems Review*, 40(4):333–344.

Weihao Zhao, Han Wu, Weidong He, Haoyang Bi, Hao Wang, Chen Zhu, Tong Xu, and Enhong Chen. 2023. Hierarchical multi-modal attention network for time-sync comment video recommendation. *IEEE Transactions on Circuits and Systems for Video Technology*.