

Short Video Segment-level User Dynamic Interests Modeling in Personalized Recommendation

Zhiyu He

hezy22@mails.tsinghua.edu.cn
DCST, Tsinghua University
Beijing, China

Zhixin Ling

lingzhixin@kuaishou.com
Kuaishou Technology
Beijing, China

Jiayu Li

DCST, Tsinghua University
Beijing, China
lijiaayu997@gmail.com

Zhiqiang Guo

georgeguo.gzq.cn@gmail.com
Kuaishou Technology
Beijing, China

Weizhi Ma

mawz@tsinghua.edu.cn
Kuaishou Technology
Beijing, China

Xinchen Luo

luoxinchen@kuaishou.com
Kuaishou Technology
Beijing, China

Min Zhang*

z-m@tsinghua.edu.cn
DCST, Tsinghua University
Beijing, China

Guorui Zhou

zhouguorui@kuaishou.com
Kuaishou Technology
Beijing, China

ABSTRACT

The rapid growth of short videos has necessitated effective recommender systems to match users with content tailored to their evolving preferences. Current video recommendation models primarily treat each video as a whole, overlooking the dynamic nature of user preferences with specific video segments. In contrast, our research focuses on segment-level user interest modeling, which is crucial for understanding how users' preferences evolve during video browsing. To capture users' dynamic segment interests, we propose an innovative model that integrates a hybrid representation module, a multi-modal user-video encoder, and a segment interest decoder. Our model addresses the challenges of capturing dynamic interest patterns, missing segment-level labels, and fusing different modalities, achieving precise segment-level interest prediction.

We present two downstream tasks to evaluate the effectiveness of our segment interest modeling approach: video-skip prediction and short video recommendation. Our experiments on real-world short video datasets with diverse modalities show promising results on both tasks. It demonstrates that segment-level interest modeling brings a deep understanding of user engagement and enhances video recommendations. We also release a unique dataset that includes segment-level video data and diverse user behaviors, enabling further research in segment-level interest modeling. This work pioneers a novel perspective on understanding user segment-level preference, offering the potential for more personalized and engaging short video experiences.

CCS CONCEPTS

• Information systems → Personalization.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '25, July 13-18, 2025, Padua, Italy

© 2024 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/xx.xxxx/xxx.xxxxx>

KEYWORDS

Video recommendation, User modeling, Segment-level interest.

ACM Reference Format:

Zhiyu He, Zhixin Ling, Jiayu Li, Zhiqiang Guo, Weizhi Ma, Xinchen Luo, Min Zhang, and Guorui Zhou. 2024. Short Video Segment-level User Dynamic Interests Modeling in Personalized Recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. ACM, New York, NY, USA, 12 pages. <https://doi.org/xx.xxxx/xxx.xxxxx>

1 INTRODUCTION

Short videos have gained immense popularity in recent years, raising the need for effective recommender systems to match users with content they may be interested in. Unlike traditional media such as movies and magazines, short videos are characterized by rapid pacing and frequent scene transitions [22, 31]. This results in highly dynamic user interests during video consumption, manifested by shifting attention and preferences across different segments. These preference shifts are often reflected in behaviors along the video timeline, such as skipping to the next video, which should be captured by models for better recommendations.

However, segment-level preferences have not been explored well in previous studies, as most existing works treat videos as whole entities, extracting features and training id embeddings to assign an overall preference score [5, 13]. When incorporating multimodal information, researchers typically focus on content features from the video cover or fixed frames [3, 17, 45], overlooking the dynamic temporal nature within videos. Preference label is typically user feedback at video-level, such as “skip” or “effective view”, while segment-level behavioral signals are severely ignored. The absence of segment-level information fundamentally limits the recommender system's accuracy. Shang et al. [36] steps further by exploring detailed visual data along the video's timeline. They constructed a graph representing user preferences implicitly by selecting positive or negative frames separately based on predefined rules. However, in real-world scenarios, video segments are tightly coupled instead of being isolated entities, as the preceding and following segments influence the user's interest in a specific segment. The separated selection overlooks the interest's continuity, resulting in an underdeveloped user interest modeling.

As shown in the left part of Figure 1, a short video typically comprises segments with different topics and rapid transitions, and

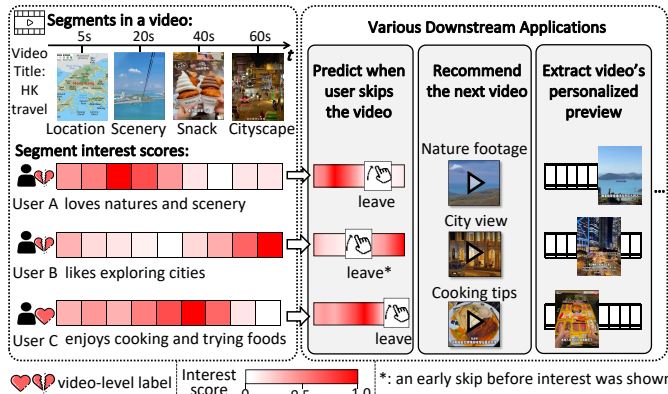


Figure 1: Users' dynamic interests in video segments reflect their diverse preferences, offering deeper insight than the overall video preference. Such interests benefit downstream applications such as video-skip prediction, recommendations, and personalized homepage thumbnails.

users' interests evolve when they engage in short video browsing. Different segments appeal to different users based on their personalized preferences. Modeling this personalized segment-level interest aware of the temporal transition, which was neglected in previous models, is crucial for short video recommender systems. Segment-level interest modeling offers significant value in practice, as illustrated in the right part of Figure 1. First is **video-skip prediction**, which identifies low-interest segments users will likely skip. It also improves **video recommendation** by incorporating segment-level interest scores, which capture deeper preferences than traditional video-level approaches. Additionally, **personalized thumbnails** in the homepage can be generated by selecting high-interest segments, driving users to click on the video and watch. These encourage a comprehensive study on segment-level interest modeling for essential user understanding, targeting considerable improvement in recommendation accuracy and user satisfaction.

Thus, we aim to model the users' interests for segments within a short video for personalized recommendations. However, modeling user segment interests is challenging: (1) User interest evolves along the video timeline, influenced by the natural temporal patterns of human attention, which are independent of video content. Effectively incorporating these patterns to capture interest dynamics is a key challenge. (2) Segment-level user feedback is typically implicit and sparse, such as scrolling actions [23], providing little information on user interest for each video segment. This lack of explicit labels complicates the task of modeling segment interest. (3) Modeling segment interests needs video content information, and the multi-modal fusion challenge arises when combining the complementary signals from user-item interactions and content-based information. To alleviate the challenges, we propose an innovative model to capture segment-level user interest, comprising a hybrid representation, a multi-modal user-video encoder, and a segment interest decoder. It serves inner-video segment position indices as embedding input and fits position bias for challenge 1. For challenge 2, we design an intra-video loss function to exploit implicit relationships between user feedback and segment interests, addressing feedback sparsity and guiding the training process. Our

model represents and encodes interactions within each modality and performs late-stage fusion to ensure modality-specific interests while allowing insights from all modalities to complement each other (challenge 3).

We present two downstream tasks to evaluate the effectiveness of our interest modeling in the recommendation. The first task is *video-skip prediction*, where we predict when or which segments a user is likely to skip based on their fine-grained interests. The second task is *video recommendation*, which integrates segment-level interest predictions to enhance the overall performance of video recommendations. The experimental results on both tasks demonstrate the accuracy of segment interest modeling and its value in recommendation. We present cases to highlight the effectiveness and potential value of interest modeling, providing a valuable perspective on segment interest modeling in recommendations. We also release a *video recommendation dataset* that includes segment-level video data and diverse user behaviors, which is the first dataset to provide both of these critical components.

The contributions of this work are as follows:

- To the best of our knowledge, we are the first to address the problem of segment-level user interest modeling in short video recommendation, extending the existing paradigm from video-level to segment-level modeling. This helps better user understanding and provides a more personalized experience.
- We propose a novel segment-level user interest model with hybrid representation, multi-modal user-video encoder, and segment interest decoder. It tackles the challenges of capturing dynamic interest, modality gaps, and missing direct segment-level user feedback.
- Experiments on a public dataset and a new dataset show encouraging results of the new paradigm and model on two tasks: video-skip prediction and video recommendation. We also release the dataset with segment-level video data and diverse user behavior data, which is the first data source for segment-level user preference understanding¹.

2 RELATED WORK

2.1 Fine-grained User Modeling

Our work shifts the focus from holistic video modeling to fine-grained segment modeling, with a deeper exploration of fine-grained user modeling. These related works can be categorized into two types: multi-dimensional interest and temporal unit-level interest.

Multi-dimensional interest breaks down user interest in an item into multiple categories or hierarchical levels. Some studies cluster items based on user interaction history and model interests across different categories [9, 16, 39]. Others examine varying behavioral feedback signals, such as clicks, purchases, and shares, to capture user preferences across different types of interactions [29]. These methods often combine coarse-grained and fine-grained modeling to improve recommendation accuracy.

Temporal unit-level interest focuses on modeling user interests based on fine-grained unit level. These researches mainly target news recommendations, where users' interest is represented by their engagement with individual tokens within an article, such as specific words or phrases. By combining token-level interest with

¹Codes and data are available at <https://anonymous.4open.science/r/SegInterest-632D/>

document-level interest, these models aim to improve the relevance of news recommendations [15, 32]. Similar to the second approach, we treat segments as the basic unit in our model. However, while news recommendation systems utilize user clicks on specific words as feedback, our work focuses on short video recommendations, where user feedback is based on the exit from segments, bringing a unique challenge to our work.

2.2 Multi-Modal Video Recommendation

Our proposed model utilizes both multimedia content and user interaction data for fine-grained segment-level user interest modeling, making it a multi-modal system. Generally, multi-modal recommender systems adopt fusion-based or graph-based methods for handling multi-modal data [26]. Fusion-based methods combine signals into unified representations using techniques like attention or bilinear pooling. For example, UVCAN [27] employs self-attention to merge user-side and item-side data, while M3CSR [3] integrates features from multiple modalities for cold-start recommendations. Graph-based methods, such as MMGCN [45] and MMGCL [46], model user-item interactions with bipartite graphs, and MMKGv [25] uses a graph attention network for knowledge graph-based fusion. These approaches aim to capture complex relationships between users and items.

However, most existing works treat videos as a whole, focusing on video-level features for recommendations. Only a few address the inherent diversity within a video. For instance, Shang et al. [36] models positive and negative frames for video-level recommendations but overlooks temporal continuity between segments. In contrast, our approach explicitly models segment-level user interest, taking into account the temporal relationships between segments. Besides, Chen et al. [6] focus on key-frame recommendations based on time-synchronized user feedback. However, our work examines user interest across different segments of a video involving downstream applications. We propose a novel model to explore how users' interests vary across different segments in short video recommendations, as well as the broader downstream applications of video-skip prediction and video-level recommendations.

2.3 Video Highlight Detection

Video highlight detection (VHD) is a topic in Computer Vision related to our work, as both divide video into segments and aim to capture content patterns in temporal sequences. The primary objective of VHD is retrieving a subset of video frames that capture a person's primary attention from the original video [34, 44]. PHD-GIFs [11] is the first personalized VHD technique that extracts highlight predictions guided by subjects' manually created GIFs. Models are constructed by convolutional layers, fully connected layers, and learned parameters based on individuals' preferred clips to guide a content-based highlight detection network [4, 34]. Bhattacharya et al. [1] builds upon this approach by employing a multi-head attention mechanism, achieving improved performance.

However, modeling segment-level user interests has fundamental differences from VHD. VHD only uses image features without user behavior as model input, while interest modeling is based on collaborative information, in which user interaction is combined with videos for personalized and comprehensive understanding.

Besides, user-uninvolved human-selected labels (GIFs) form the datasets for VHD training and evaluation, while in real-world short video recommendation scenarios, it's difficult to explicitly figure out users' "chosen segments", which is discussed in Section 1 as challenge 2. As a result, further discussion or comparison between VHD and our proposed model is biased and not objective, thus is not included in the remainder of this paper.

3 USER SEGMENT INTEREST MODELING

3.1 Problem Definition

In short video platforms, users continuously watch videos ranging from a few seconds to a few minutes. Videos are presented one at a time, and users watch the next system-recommended video by swiping up or down the screen. Formally, $\mathcal{U} = \{u_1, u_2, \dots\}$ represents the set of all users and $\mathcal{V} = \{v_1, v_2, \dots\}$ represents the set of all items (i.e., videos).

Video segments are defined by evenly dividing videos with the same time intervals: Each video $v \in \mathcal{V}$ is divided into N segments, $S_v = (S_v^1, S_v^2, \dots, S_v^N)$, where the duration of S_v^i is t for any i . The viewed segments are determined with the viewing time of each video, and we assume that users watch videos from the beginning, as this behavior is common in video streaming.

We define the **segment interest modeling** problem in the context of sequential recommendation, where the user's history of viewed videos is denoted as (v_1, v_2, \dots) . The corresponding user history of u viewed segments is denoted as $S_u = (S_u^1, S_u^2, \dots, S_u^M)$, where M is the number of viewed segments. Given a user u and a target video v , our goal is to infer the interest score sequences $\vec{p} = (p_1, p_1, \dots, p_N)$ for video segments S_v , where p_i is the interest score for segment S_v^i . These segment interest scores can be used for downstream applications.

3.2 Model Overview

We propose a model for user segment interest with Hybrid User & video Representation, Multi-modal User-video Encoder, Segment Interest Decoder, as shown in Figure 2.

Hybrid User-video Representation learns embeddings from different modalities for the user and target video. The ID modality is represented through embedding layers, while feature encoders process other modal inputs, such as visual features. The representations from the user and video interact independently for each modality in the Multi-modal User-video Encoder. We design modal-aware interest detection with user-video cross-attention. It outputs modal-specific interest scores from the interaction information of each modality. Finally, the Segment Interest Decoder fuses different modalities' outputs and generates segment interest scores.

Notably, the model is modality-agnostic; it can easily accommodate changes in the number or type of modalities.

3.3 Hybrid User & video Representation

User and video representations are derived from inputs across different modalities. This work represents users and videos using two modalities: the visual modality and the ID modality. The visual modality captures content-related features, and this representation process can be extended to other modalities, such as auditory and

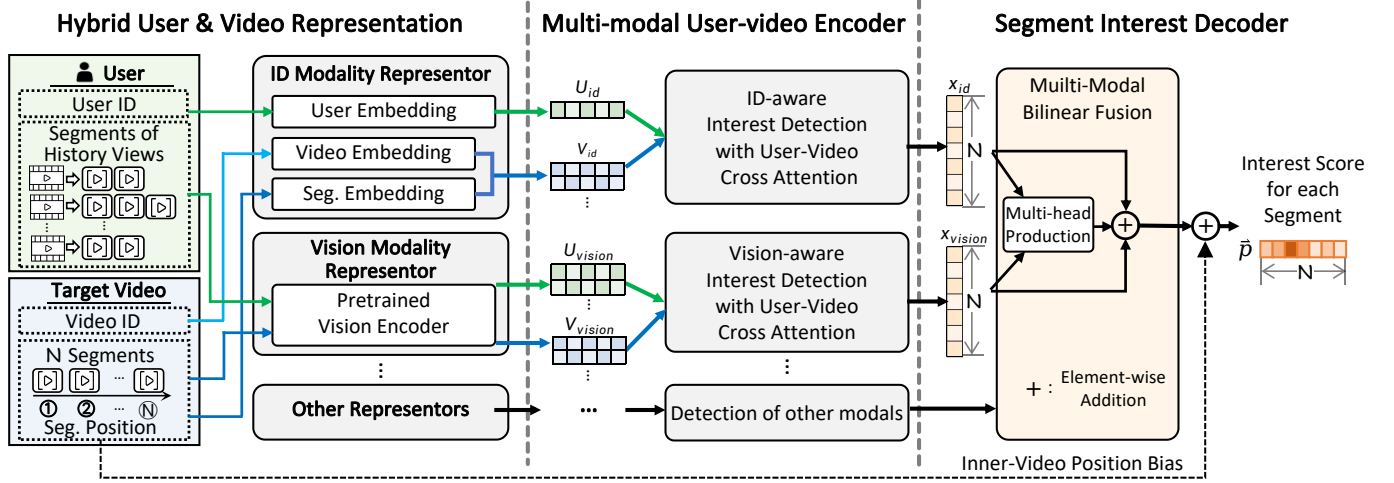


Figure 2: Overview of user segment interest modeling with hybrid user and video representation, multi-modal user-video encoder, and segment interest decoder. N is the number of segments in the target video, and segment interest scores are the model's output.

textual. The ID modality, on the other hand, consists of user and video identifiers, enabling personalization.

As for the visual modality, we leverage CLIP² [33] as the pre-trained Vision Encoder to extract feature representations. These features are then mapped to the desired representation space through a linear projector, resulting in U_{vision} and V_{vision} , derived from the user's viewed segments S_u and target video segments S_v inputs, respectively. For ID modality, we first embed the user ID to obtain the user embedding U_{id} . To distinguish between different segments within the same video, we introduce segment positions ranging from 1 to N . We then embed the video ID and the segment positions. The video ID's embedding and segment position embeddings are concatenated to form the video segment representations, V_{id} .

This design strikes a balance between content-driven semantics and unique identifiers, allowing the model to effectively capture both contextual and personalized information.

3.4 Multi-modal User-video Encoder

The Multi-modal User-video Encoder includes Modal-aware Interest Detection to achieve segment-related interest representations by processing user-video interactions within each modality separately.

Figure 3 demonstrates Modal-aware Interest Detection. The core module of Modal-aware Interest Detection is User-Video Cross-attention. Unlike conventional transformer-based methods that treat users and items uniformly within a shared sequence, we recognize that user and item representations possess distinct syntactic and semantic structures. Inspired by Wang et al. [42], we design the User-Video Cross-Attention. For simplicity, we denote U_{vision} and U_{id} as U , and V_{vision} and V_{id} as V . In the vision modality, $V \in \mathbb{R}^{N \times d}$ and $U \in \mathbb{R}^{M \times d}$, while in ID modality, $U \in \mathbb{R}^{1 \times d}$. Due to the use of the user's historical sequence, the number of segments viewed by the user is greater than the number of segments in the target video ($M > N$).

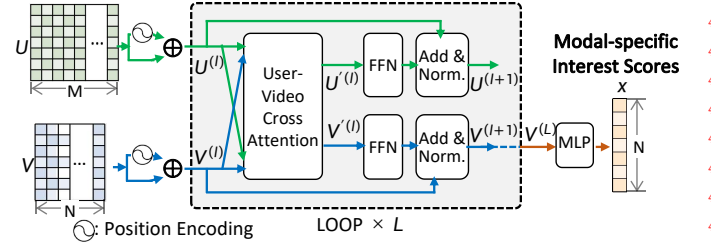


Figure 3: Details of the interest detection module in the multi-modal encoder. U and V denote the representations of the user and target video.

We first encode sequence positions to incorporate the order of elements. Then, the model enters a loop with each layer contains a User-video Cross-attention. The User-video Cross-attention aggregates the user embeddings $U^{(l)}$ and video embeddings $V^{(l)}$ to form user-interacted video representations, $V'^{(l)}$, through the following process:

$$A_{VU}^{(l)} = \text{FFN}(V^{(l)}) \cdot \text{FFN}(U^{(l)}) / \sqrt{d} \quad (1)$$

$$A_{VV}^{(l)} = \text{FFN}(V^{(l)}) \cdot \text{FFN}(V^{(l)}) / \sqrt{d} \quad (2)$$

$$V'^{(l)} = \text{Softmax}(A_{VU}^{(l)} \oplus A_{VV}^{(l)}) \cdot (\text{FFN}(V^{(l)}) \oplus \text{FFN}(U^{(l)})) \quad (3)$$

where \oplus represents the concatenation operation. A similar set of equations is used to obtain $U'^{(l)}$:

$$A_{UU}^{(l)} = \text{FFN}(U^{(l)}) \cdot \text{FFN}(U^{(l)}) / \sqrt{d} \quad (4)$$

$$U'^{(l)} = \text{Softmax}(A_{VU}^{(l)} \oplus A_{UU}^{(l)}) \cdot (\text{FFN}(V^{(l)}) \oplus \text{FFN}(U^{(l)})) \quad (5)$$

Through Feedforward Network (FFN), residual connections, and layer normalization, we obtain $U^{(l+1)}$ and $V^{(l+1)}$ from $U'^{(l)}$ and

²<https://huggingface.co/openai/clip-vit-large-patch14-336>

$V^{(L)}$, U^{L+1} and V^{L+1} are inputs to the next layer. We adopt L layers to aggregate the user embedding. The encoder's output, $V^{(L)}$, is taken as video segment-related latent scores, aggregated from user representations. Subsequently, we employ an MLP (Multilayer Perceptron) module to reduce the representations from high to low dimensions, then we have segment interest scores x from different modalities.

3.5 Segment Interest Decoder

The segment interest decoder receives the latent scores of the user-interacted video segments to decode them into the final interest score. Since those scores result from various modalities, multi-modal bilinear fusion is adapted to preserve their separate strengths while enabling complement integration among each other as:

$$o = \sigma \left(b_f + \sum_i \text{Proj}(x_i) + \sum_{i \neq j} \text{Proj}(x_i^\top x_j) \right) \quad (6)$$

where x_i represents the output from different Modal-aware Interest Detection, $\text{Proj}(\cdot)$ denotes independently-trained Linear Projectors and b_f is a trainable bias term. $x_i^\top x_j$ refers to the multi-head product of the modality-specific features, which enables capturing diverse interactions while enhancing computational efficiency.

To model the human inherent temporal attention pattern, for example, user interest typically accumulates or decreases as they watch a video from start to finish, and the probability of interest decreases for later segments, we introduce an inner-video position bias to capture the relationship between segment position and interest. This position bias is added to the fusion output o , producing the final segment interest score \vec{p} :

$$\vec{p} = o + (w_p \cdot \vec{pos} + b_p) \quad (7)$$

where w_p and b_p are learnable parameters for position bias, $\vec{pos} = (idx_1, idx_2, \dots, idx_N)$. $\vec{p} = (p_1, p_1, \dots, p_N)$ represents the interest scores for the segments, with p_i being the interest score for segment S_v^i .

3.6 Training and Inference

As discussed in Section 1, it's unaffordable to have user feedback for every video segment. In real-world applications, the only practical user interaction is the skipped segment position, during which he/she gets bored or distracted and skips to the next video. Since segment interests are reflected in the differences between segments within the same video, a carefully designed intra-video loss function based on the skipped segment position is proposed to fully exploit the limited user interaction within video watching. Specifically, the loss function assumes that the skipped segment position should have the lowest interest level among the watched segments, with the loss for one video segment pair (u, v) :

$$\mathcal{L}_{u,v} = - \sum_{(i=y, j \neq y)}^{N-1} \ln \sigma(p_j - p_i) \quad (8)$$

where interest score p_i indicates the likelihood of continuing watching for the user, y denotes the skipping segments of user u toward target video v , and N is the segment number.

The overall loss is defined as the average of $\mathcal{L}_{u,v}$ across user-video interactions in the training set:

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{(u,v) \in \mathcal{D}} \mathcal{L}_{u,v} \quad (9)$$

During inference, the model directly computes the interest scores for each segment in the target video.

3.7 Potential Application for Downstream Tasks

The segment interest scores obtained from our model encapsulate rich information about both user preferences and video content, closely tied to recommendation systems. These scores open up various downstream applications, such as enhancing video-level recommendation performance, video automatical editing and thumbnail generation. To illustrate the effectiveness of our proposed user interest scores, we conduct two tasks: (1) Video-skip prediction (Section 4) to directly validate the accuracy of user interest scores, and (2) Short video recommendation (Section 5) as a common and important task to evaluate whether segment interest scores are useful in real-world scenarios.

4 VIDEO-SKIP PREDICTION (TASK 1)

4.1 Task Description and Settings

4.1.1 Task Description. The video-skip prediction task predicts in which segment users will possibly skip the video. It attempts to deeply characterize user interest and video information in the recommender system, whose results can serve as a supplementary metric in online recommendations. After the generation of segment-level interest scores (outputs of Section 3), segments are ranked as the scores are negatively correlated to the likelihood of skipping. For each user-video interaction, we assess whether the actual skipped segment appears in the top-K segments (HR@K) and compute the corresponding Normalized Discounted Cumulative Gain (N@K). We set $K=1,5,10$.

4.1.2 Dataset. Based on our task objectives, the short video recommendation dataset we used must include user timestamp feedback on videos (watch time or skipping). Additionally, it is preferable to have segment-level modality features (visual, auditory, and textual) to capture video content. Currently, no public dataset contains both types of information. Here, we release a dataset (SegMM) to fill this gap. Additionally, some datasets include ID and behavioral features. Considering the differences in time span and the number of interactions compared to SegMM, we select the KuaiRand dataset. Here's the description of two datasets:

SegMM is a dataset from a commercial short video platform. It contains 2,369 randomly sampled active users with millions of interactions over 3 days (June 01 to June 03, 2024). It has vision features for each segment of raw videos and the behaviors of these users' interaction. To the best of our knowledge, it's the first short video recommendation dataset with both fine-grained features and behaviors. *We will publish this dataset with the work.*

KuaiRand [10] is collected from one popular Chinese short video app, KuaiShou. We use the data of recommended items. It contains 983 users and their interactions with system-recommended short videos over two weeks (April 07 to April 21, 2022).

Note that we considered MicroLens [30] as it's a valuable dataset with raw videos; however, it lacks user timeline behavior data.

Table 1: Datasets statistics. “Visual” refers to whether the data provides visual features.

Datasets	#Users	#Items	#View	#Segment	Visual
SegMM	2,369	362,430	902,115	3,920,483	✓
KuaiRand	983	717,652	1,615,315	8,140,477	–

Interactions with fast-forwarding operations are excluded in SegMM. Therefore, we can assume that videos are displayed from the beginning, and the skip position of each video is directly derived from users’ watch time. We partitioned the data using an 8:1:1 user-based split. Short videos typically range from a few seconds to several minutes. To facilitate task evaluation, we selected videos no longer than 200 seconds and assumed each segment to be 5 seconds. The statistics of datasets are shown in Table 1.

4.1.3 Baselines. Three types of baselines are employed in this task. Intuitively, the most popular method of calculating position probabilities is applied. Besides, since few prior works concerned the video skip prediction task, three traditional recommenders are implemented by generating prediction scores for each segment and ranking those scores inside videos for final evaluation. Furthermore, watch-time prediction is also compared due to its similarity to the target task. Follows detail those baselines:

- (1) **MostPopular** uses skip probabilities for each segment position, derived from statistics on the training and validation datasets. It includes random selection, overall position probability (AllPosition), user-specific position probability (UserPosition), and item-specific position probability (ItemPosition).
- (2) **GeneralRec** represents standard recommendation methods employing collaborative filtering techniques such as LightGCN [14] and DirectAU [40]. Additionally, it includes sequence-based recommenders like Caser [37] and SASRec [18].
- (3) **ContextRec** incorporates contextual information into recommendations. We include classic models like WideDeep [7] and DCNv2 [43], as well as state-of-the-art methods such as FinalMLP [28] and AdaGIN [35]. It also considers historical context with models like DIN [50], DIEN [49], and CAN [2].
- (4) **MMRec** encompasses multi-modal recommendation methods that integrate various data modalities, including SLMRec [38], BM3 [53], and FREEDOM [52].
- (5) **Watch-time Prediction Methods** including WLR [8] by positive sample weighting, D2Q [47] by group regression, and TPM [24] as a representative work of and tree-based classification methods.

4.1.4 Implementation Details. We implement recommender methods using ReChorus [20, 41] and multi-modal recommender using MMRec [51]. The baselines of the Watch time prediction follow the original paper’s code. Similar to the TPM method, proportion normalization of the interest scores’ reciprocal is used as probability weight as the result of watch time when computing MAE. We implement our method with PyTorch³. The model is optimized by Adam [19] optimizer with tunable learning rate and embedding size, where the batch size is 1024. Learning rate are tuned from 1e-4 to 1e-2 and embedding size are from 32 to 128. As for our model and the baselines that consider history, we set max history as 20. We

³<https://pytorch.org/>

use NDCG@5 on the valid set for early stopping if the performance does not increase in 10 epochs.

4.2 Overall Video-skip Ranking Performance

Table 2 and Table 3 show the performances of our proposed method compared to baselines. Our proposed method demonstrates significant improvements over baselines across two datasets, thereby validating the accuracy of segment interest modeling. We highlight its effectiveness in capturing user interests at the segment level to boost video-skip prediction.

Analyzing the performance of different baseline categories, we find that segment interests are notably influenced by position within the video. Among the most popular strategies, item-specific position probability yields superior performance, particularly in the HR@1 metric. This suggests a high consistency in user skipping behavior within the same video. Some recommendation methods, such as LightGCN, perform even worse than the most popular category since they are not specified for segment interest and cannot characterize the segment information. When evaluating segment recommendation scores and interest rankings, AdaGIN emerges as the top-performing baseline for the SegMM dataset, while CAN leads for the KuaiRand dataset, indicating that context-aware methods hold a distinct advantage in this task. For the MMRec category, BM3 and FREEDOM perform better than their backbone (LightGCN) by including visual features.

For Table 3, the performance comparison highlights the ability of our model to estimate watch time.

In summary, our method achieves remarkable performance in video-skip prediction, a critical subtask within recommendation systems. By accurately modeling segment-level user interests, our approach effectively identifies points of lowest user engagement where skips are most likely to occur. The experimental results underscore the superiority of our method, ensuring more reliable and personalized video recommendations.

4.3 Video-skip Ranking Performance on Cold-start Items

Utilizing item-specific position probability (ItemPosition) significantly outperforming the global position probability (AllPosition), demonstrating consistency in the skipping positions in the same video. It raises concerns about the performance of cold videos. Approximately one-third of the general test set is with cold video, and their performance is shown in Table 4.

The experimental results indicate that the recommender methods consistently outperform the most popular approach for cold items, and our method still significantly outperforms the baselines on both cold and non-cold items, reflecting the model’s generalization capability in understanding videos. Incorporating segment position information into our model ensures that it is not entirely unaware of cold videos. Comparing Table 4 and Table 2, our proposed model performs slightly worse on cold videos when only using ID input. The performance on cold videos improves by incorporating content information (Visual modality), demonstrating its generalization capability in understanding videos.

Table 2: The video-skip prediction performance. * and ** indicate p-value<0.1 and <0.05 from t-test between *Ours* and the best baseline. Bold and underline show the best of all results and the best baseline results. N is short for NDCG.

Category	Dataset	SegMM					KuaiRand				
	Methods	HR@1	HR@5	N@5	HR@10	N@10	HR@1	HR@5	N@5	HR@10	N@10
MostPopular	Random	0.0244	0.1258	0.0737	0.2530	0.1142	0.0251	0.1266	0.0746	0.2515	0.1145
	AllPosition	0.1309	0.2625	0.2007	0.3679	0.2343	0.0897	0.2111	0.1523	0.3242	0.1884
	UserPosition	0.1504	0.2823	0.2204	0.3850	0.2531	0.0992	0.2219	0.1629	0.3336	0.1985
	ItemPosition	<u>0.2619</u>	0.3824	0.3269	0.4701	0.3548	<u>0.2648</u>	0.3790	0.3299	0.4669	0.3580
GeneralRec	LightGCN	0.1148	0.2380	0.1808	0.2734	0.1924	0.0866	0.2199	0.1567	0.2803	0.1762
	DirectAU	0.1684	0.3680	0.2770	0.4128	0.2917	0.1263	0.3284	0.2334	0.4065	0.2587
	Caser	0.1655	0.3742	0.2782	0.4296	0.2963	0.1216	0.3091	0.2207	0.3816	0.2443
	SASRec	0.1831	0.4381	0.3201	0.5048	0.3420	0.1554	0.3940	0.2821	0.4839	0.3112
ContextRec	WideDeep	0.1306	0.2765	0.2095	0.3118	0.2211	0.1231	0.3404	0.2381	0.4311	0.2675
	DCNv2	0.1525	0.3518	0.2605	0.4054	0.2780	0.1420	0.4419	0.2997	0.5665	0.3401
	FinalMLP	0.1663	0.4032	0.2941	0.4683	0.3154	0.1465	0.4336	0.2979	0.5571	0.3379
	AdaGIN	0.2177	<u>0.5868</u>	<u>0.4167</u>	<u>0.6778</u>	<u>0.4465</u>	0.1597	0.4376	0.3066	0.5539	0.3442
	DIN	0.1312	0.4030	0.2744	0.4988	0.3055	0.1542	0.4177	0.2934	0.5259	0.3286
	DIEN	0.1622	0.4118	0.2955	0.4920	0.3217	0.1812	<u>0.4535</u>	<u>0.3247</u>	<u>0.5702</u>	<u>0.3625</u>
	CAN	0.2031	0.4817	0.3517	0.5640	0.3785	0.1762	0.4280	0.3095	0.5302	0.3426
MMRec	SLMRec	0.1136	0.2333	0.1773	0.2745	0.1908	(no visual data)				
	BM3	0.1506	0.3639	0.2641	0.4387	0.2886					
	FREEDOM	0.1808	0.3783	0.2874	0.4243	0.3025					
Ours	ID	0.3353**	0.7676**	0.5462**	0.8534**	0.5740**	0.2904	0.5709	0.4275	0.7378	0.4818
	Visual	0.3828**	0.8171**	0.6186**	0.8567**	0.6318**	(no visual data)				
	Both	0.4072**	0.8214**	0.6228**	0.9225**	0.6572**	0.2904*	0.5709**	0.4275**	0.7378**	0.4818**

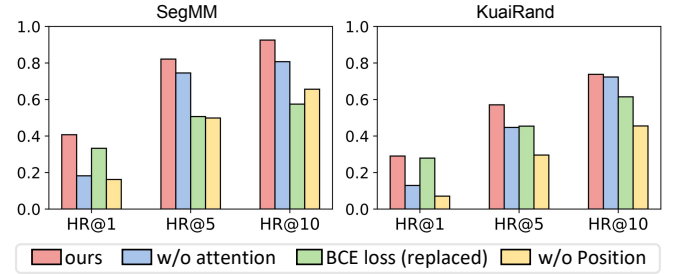
Table 3: The performance comparison between our model and baselines on the watch-time prediction.

Dataset	Metrics	WLR	D2Q	TPM	Ours		
					ID	Visual	Both
SegMM	HR@1	0.1436	0.1227	<u>0.3049</u>	0.3353**	0.3828**	0.4072**
	MAE ↓	3.3790	<u>3.1228</u>	3.3217	3.0254*	3.0399*	2.9564**
KuaiRand	HR@1	0.0564	0.1321	<u>0.1663</u>	0.2904** (no visual data)		
	MAE ↓	4.5244	<u>3.9337</u>	4.2714			

MAE (Mean Absolute Error) is a typical measure of regression accuracy in watch time prediction tasks.

Table 4: Video-skip prediction performance on cold videos.

Dataset	SegMM			KuaiRand		
	H@1	H@5	H@10	H@1	H@5	H@10
Methods						
AllPosition	0.1395	0.2764	0.3795	0.0887	0.2109	0.3229
UserPosition	0.1672	0.3001	0.3959	0.1040	0.2291	0.3382
DirectAU	0.1649	0.4015	0.4639	0.1302	0.3587	0.4454
SASRec	0.1916	0.4553	0.5334	0.1790	0.4169	0.5023
AdaGIN	<u>0.2045</u>	<u>0.5621</u>	<u>0.6642</u>	0.1623	0.4496	0.5693
DIEN	0.1719	0.4586	0.5493	<u>0.1940</u>	<u>0.4693</u>	<u>0.5700</u>
FREEDOM	0.1812	0.3801	0.4220	(no visual data)		
Ours(ID)	0.3211	0.7638	0.8496	0.2662	0.5582	0.7308
Ours(Visual)	0.4060	0.8073	0.8519	(no visual data)		
Ours(Both)	0.4068	0.8170	0.9234	0.2662	0.5582	0.7308

**Figure 4: Ablation study of attention mechanism, segment position indices, and replacing our loss with BCE (Binary Cross-Entropy) loss.**

4.4 Ablation Study for Video-skip Prediction

We conduct ablation studies on the input modality, interest detection module, segment position, and intra-video loss.

Our method fuses multiple modalities, combining collaborative (ID) and visual information. As shown in Table 2, the visual modality outperforms the ID modality. This is because the visual modality leverages both user history and target video content, with the encoder effectively capturing their relationships. Bilinear fusion of modalities yields the best performance, highlighting the complementary nature of different features.

We also remove cross-attention in the interest detection module and segment position indices. The attention module captures user-video relationships, while segment position indices are integrated into segment embeddings and intra-video position bias. Figure 4 shows performance dropped when they were removed, highlighting the importance of user-video cross-attention and segment position

for the model. Especially for KuaiRand, where there's no visual data, segment positions are crucial for distinguishing between segments.

Finally, we replaced our intra-video loss with BCE (binary cross-entropy) loss. BCE treats each segment independently and assumes skipped positions have an interest label of 0, causing a significant performance drop. This highlights the effectiveness of our method in addressing challenges with missing segment-level labels.

5 SHORT VIDEO RECOMMENDATION(TASK 2)

5.1 Segment-integrated Video Recommendation Framework

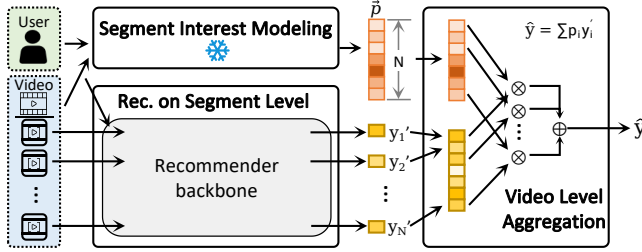


Figure 5: SegRec: Segment-integrated Video Recommendation Framework. The parameter of segment interest modeling is frozen, serving segment interest scores $\vec{p} = (p_1, p_2, \dots, p_N)$ to down-streaming video recommendation.

While segment interest modeling plays a key role in video-skip prediction, it also provides valuable insights that can enhance video-level recommendation by offering a deeper understanding of user preferences. We conduct video-level recommendations to evaluate the effectiveness and usefulness of segment interest modeling. Therefore, segment interest scores are introduced intuitively to existing recommendation models. We propose SegRec, a framework that integrates segment interest for video-level recommendations.

Figure 5 illustrates **SegRec**, a segment-integrated video recommendation framework. For each user and target video pair, given the user's historical interactions, we compute segment interest scores p_i (as described in Section 3), which represent the importance of each segment in recommending the video. The same user and video features are input into the backbone recommendation model, which predicts scores y_i for each segment. These scores are then weighted and aggregated at the video level to obtain the final prediction score $\hat{y} = \sum_{i=1}^L \hat{p}_i y_i$, where $\hat{p} = \text{Softmax}(\vec{p})$.

Following standard practice, we apply binary cross-entropy loss between the predicted score \hat{y} and the label y , which optimizes the parameters of the backbone model.

5.2 Task Settings

5.2.1 Backbones and Baseline Settings. We integrated SegRec on several state-of-the-art context-aware methods as backbones, including interaction-based and sequential ones. The context-aware recommenders include WideDeep [7] and AdaGIN [35], which are from the classic to the latest. By adding history information, context-sequential-aware recommenders include DIN [50] and CAN [2].

As for baseline setting, we use self-information recommend without segments (Video), segment-level integration by sum (SegSum) and by learnable weight (SegAdjust). Besides, we serve segment

interest scores from task 1, including ItemPosition, the baseline model that is the same as the backbone, and the best-performing baseline from each category as representatives.

5.2.2 Implementation Details. Following previous work [12, 21, 48], we define video-level labels for the Click-Through Rate (CTR) prediction task with “effective-view” as follows: whether watch time exceeds the threshold of the corresponding duration bucket. In practice, we divide the video duration into 10 buckets and use the median of the view ratio within each bucket as the label threshold. To prevent information leakage, the data split strategy follows the same approach as in Task 1. Notably, all baselines incorporate video duration and multimodal features as video-side features, with multi-modal features mapped to a lower-dimensional space. Since directly introducing multi-modal features does not always improve performance, we select the best-performing combination of features to ensure fairness. The model hyperparameter tuning ranges are consistent across both the baseline and our SegRec.

5.3 Video Recommendation Performance

Table 5 shows the recommendation results on SegMM and KuaiRand.

The experiments demonstrate that our method outperforms the baseline models across different backbones, with significant gains with WideDeep and DIN as backbones. Direct video recommendation performs well, as does SegAdjust, which benefits from trained weights for each segment score. In comparison, our approach leverages segment interest scores, which serve as valuable weights for individual segments, leading to targeted improvements in overall model performance. The consistent enhancement underscores the utility and effectiveness of segment interest modeling.

When applying the output of Task 1's baseline as the weight, we observed that their performance generally aligns with its performance in Task 1. For example, AdaGIN, as the best baseline for SegMM in Task 1, is also the best baseline when used for video recommendations. However, not all segment interests contribute positively to the video recommendation task, suggesting that some segment interests may introduce noise into recommendation. In contrast, our segment-level interest modeling always outperforms, further demonstrating the effectiveness of accurate interest scores. Regarding modality fusion, combining modalities (both) yields the best results, with ID consistently outperforming visual modality.

To be noted, our framework is designed primarily to validate the effectiveness of segment interest scores. We find the utility of these scores in improving performance while maintaining a directed and explainable approach. This provides insight into how segment-level data can be served effectively and interpretable without overcomplicating the recommendation process.

6 CASE STUDY AND DISCUSSIONS

In previous sections, we proposed segment-level interest modeling and evaluated it on two typical tasks with improved accuracy. As shown in Figure 6, several cases visually present the model's performance, highlighting our predicted segment interest based on segment images, position indices, and user history. Each case in the test set presents segment images in sequence. The predicted interest scores, derived from the model's inference through proportional

Table 5: The video-level recommendation performance. Bold and underline show the best of all results and the best baseline results. */ indicates p-value<0.1/0.05 from the t-test conducted over five repetitions.**

SegMM													
	Backbone	WideDeep			AdaGIN			DIN			CAN		
Group	Methods	AUC	F1	Logloss↓	AUC	F1	Logloss↓	AUC	F1	Logloss↓	AUC	F1	Logloss↓
Self-Info	Video	0.7301	0.6550	0.6188	0.7371	0.6634	0.6032	0.7471	0.6809	0.5992	0.7502	0.6901	0.5927
	SegSum	0.7310	0.6597	0.6185	0.7347	0.6669	0.6078	0.7438	0.6748	0.6068	0.7495	0.6861	0.5931
	SegAdjust	0.7312	<u>0.6682</u>	0.6154	0.7382	0.6709	0.5978	<u>0.7489</u>	0.6808	0.6006	<u>0.7539</u>	<u>0.6926</u>	<u>0.5887</u>
Weight from Task1 baseline [†]	ItemPosition	0.7329	0.6588	0.6163	0.7418	0.6671	0.6082	0.7433	0.6680	0.5998	0.7423	0.6392	0.6048
	Backbone	0.7295	0.6517	0.6151	<u>0.7421</u>	<u>0.6833</u>	<u>0.5968</u>	0.7426	0.6602	0.6009	0.7388	0.6901	0.6092
	SASRec	0.7288	0.6503	0.6244	0.7298	0.6689	0.6182	0.7358	0.6623	0.6074	0.7419	0.6595	0.6022
	AdaGIN	<u>0.7338</u>	0.6612	<u>0.6085</u>	<u>0.7421</u>	<u>0.6833</u>	<u>0.5968</u>	0.7464	<u>0.6812</u>	<u>0.5971</u>	0.7494	0.6890	0.5932
	FREEDOM	0.7322	0.6568	0.6108	0.7412	0.6720	0.6003	0.7409	0.6799	0.5995	0.7445	0.6852	0.5975
SegRec	ID	0.7439**	0.6742**	0.5963**	0.7443*	0.6883*	0.5957	0.7584**	0.6910**	0.5834**	0.7579*	0.6949	0.5846*
	Vision	0.7419**	0.6725*	0.5970*	0.7424	0.6853	0.5972	0.7580**	0.6904**	0.5837**	0.7557	0.6921	0.5865
	Both	0.7441**	0.6762**	0.5959**	0.7452*	0.6919*	0.5945	0.7581**	0.6915**	0.5827**	0.7558	0.6997	0.5875

KuaiRand													
	Backbone	WideDeep			AdaGIN			DIN			CAN		
Category	Methods	AUC	F1	Logloss↓	AUC	F1	Logloss↓	AUC	F1	Logloss↓	AUC	F1	Logloss↓
Self-Info	Video	<u>0.7214</u>	<u>0.6449</u>	<u>0.6133</u>	0.7313	0.6616	0.6059	0.7347	0.6697	0.6039	0.7450	0.6752	0.5953
	SegSum	0.7027	0.6383	0.6363	0.7301	0.6606	0.6071	0.7283	0.6677	0.6086	0.7447	0.6787	0.5951
	SegAdjust	0.7073	0.6403	0.6342	<u>0.7318</u>	<u>0.6629</u>	<u>0.6031</u>	0.7325	<u>0.6760</u>	0.6059	<u>0.7454</u>	<u>0.6792</u>	<u>0.5950</u>
Weight from Task1 baseline [†]	ItemPosition	0.6972	0.6416	0.6369	0.7253	0.6526	0.6115	0.7372	0.6563	0.6013	0.7338	0.6353	0.6105
	Backbone	0.6900	0.6303	0.6345	0.7156	0.6501	0.6176	0.7415	0.6525	0.5977	0.7399	0.6676	0.5995
Task1 baseline [†]	SASRec	0.6936	0.6232	0.6321	0.7161	0.6508	0.6244	0.7376	0.6590	0.6012	0.7369	0.6400	0.6071
	DIEN	0.6877	0.6415	0.6379	0.6986	0.6226	0.6320	<u>0.7419</u>	0.6749	<u>0.5974</u>	0.7408	0.6642	0.5990
SegRec	ID	0.7276*	0.6713**	0.6092*	0.7339*	0.6723*	0.6020	0.7445	0.6874**	0.5953	0.7462	0.6800	0.5938

[†] means baselines in video-skip prediction (Task 1). SASRec is the best baselines of GeneralRec in Task 1. AdaGIN and DIEN are the best baselines of ContextRec for SegMM and KuaiRand datasets in Task 1, respectively. We choose the FREEDOM as baseline for SegMM dataset with visual features.



Figure 6: Segment interest cases with represented images, position indices, our model-predicted interest scores (normalized), and user behaviors (unknown to the model). \rightarrow represents keep viewing and \times is video skipping.

normalization, are displayed as a heatmap, with user viewing and skipping behaviors shown below.

Case 1A (abbreviation of user A interacting with video 1, similarly for other cases) and 1B demonstrate the user-aware generalization ability with two disparate prediction-behavior pairs for

video 1. The video spends a long period (segments 2-4) on emoticons without substantive information, which is distractive for most users, and successfully captured by the model via giving a relatively low-interest score for both users. Based on personalized history views, the model outputs even lower interest scores for the emoticons for user A than user B, showing remarkable consistency with actual behaviors: skipping at segment 2 and watching the whole video, respectively. Besides, we examine the video recommendation prediction score in Task 2 for these two cases, using AdaGIN as a representative of the latest models as the backbone. **SegRec** improved the recommendation scores (closer to true labels) by incorporating segment-interest from 0.2144 to 0.1889 for case 1A (label=0) and from 0.3601 to 0.5936 for the case 1B (label=1) against baseline method without interest scores, showing the importance of segment interest scoring.

Case 2C and 3D involve cold items, showcasing the model's adaptability and ability to understand multimodal information. Case 2C is an example of scene transition, as segments 1-9 introduce the production procedure of instant noodles, and the following segments focus on cooking and cuisine. The model gives a much higher score for the latter part, inferring the user may prefer food and delicacy. However, user C skipped early at segment 2, missing possible interest parts, indicating that segment-level interest modeling can help recommendation by parts of the target instead of the

entire video. In the advertisement video 3, segments 2 and 8 display the product’s appearance, while the others introduce its usage. The model outputs a relatively low interest score at the product segments, meeting the user behavior of skipping at segment 2.

The results highlight the potential of our model in **video personalized thumbnail generation**, allowing for the display of interest thumbnails on users’ homepages to attract clicks. It enables **personalized video editing suggestions**, where uninterested sections are automatically fast-forwarded, and content is tailored to users’ specific interests. Unfortunately, those untapped tasks were not included in this work due to the lack of valid datasets. We sincerely call upon more informative datasets for future studies.

7 CONCLUSIONS

To our knowledge, this is the first research in the recommendation field focusing on segment-level dynamic user interest. We propose a novel modeling method to extend the paradigm of the recommendation system from the video level to the segment level. Consisting of hybrid representation, multi-modal encoder, and segment interest decoder, the model excelled at video skip prediction and recommendation tasks, evaluated by experiments across two datasets and demonstrated by several cases, showing its ability in personalized recommendation and potential for more complicated tasks.

We hope this work can pave the way beyond traditional content recommendation by focusing on delivering content tailored to users’ evolving interests. With the development of valid datasets, we aim to conduct more research to explore the field of segment characterization, bridging the gap between human cognition and innovative, practical video applications in the future.

REFERENCES

- [1] Uttaran Bhattacharya, Gang Wu, Stefano Petrangeli, Viswanathan Swaminathan, and Dinesh Manocha. 2022. Show Me What I Like: Detecting user-specific video highlights using content-based multi-head attention. In *Proceedings of the 30th ACM International Conference on Multimedia*. 591–600.
- [2] Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Yong-Nan Zhu, Zhangming Chan, Na Mou, et al. 2022. CAN: feature co-action network for click-through rate prediction. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 57–65.
- [3] Gaode Chen, Ruina Sun, Yuezihan Jiang, Jiangxia Cao, Qi Zhang, Jingjian Lin, Han Li, Kun Gai, and Xinghua Zhang. 2024. A Multi-modal Modeling Framework for Cold-start Short-video Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 391–400.
- [4] Runnan Chen, Penghao Zhou, Wenzhe Wang, Nenglu 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*. 7980–7989.
- [5] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. 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*. 1146–1153.
- [6] Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin. 2017. Personalized key frame recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 315–324.
- [7] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishu Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ipsir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.
- [8] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.
- [9] Yingpeng Du, Ziyang Wang, Zhu Sun, Yining Ma, Hongzhi Liu, and Jie Zhang. 2024. Disentangled Multi-interest Representation Learning for Sequential Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 677–688.
- [10] Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. 2022. KuaiRand: An Unbiased Sequential Recommendation Dataset with Randomly Exposed Videos. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3953–3957.
- [11] Ana Garcia del Molino and Michael Gygli. 2018. Phd-gifs: personalized highlight detection for automatic gif creation. In *Proceedings of the 26th ACM international conference on Multimedia*. 600–608.
- [12] Xudong Gong, Qinlin Feng, Yuan Zhang, Jiangling Qin, Weijie Ding, Biao Li, Peng Jiang, and Kun Gai. 2022. Real-time short video recommendation on mobile devices. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 3103–3112.
- [13] Pan Gu and Haiyang Hu. 2024. A holistic view on positive and negative implicit feedback for micro-video recommendation. *Knowledge-Based Systems* 284 (2024), 111299.
- [14] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.
- [15] Yun Hou, Yuanxin Ouyang, Zhuang Liu, Fujing Han, Wenge Rong, and Zhang Xiong. 2023. Multi-level and Multi-interest User Interest Modeling for News Recommendation. In *International Conference on Knowledge Science, Engineering and Management*. Springer, 200–212.
- [16] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation. In *Proceedings of the 28th ACM International conference on Multimedia*. 3487–3495.
- [17] Peiguang Jing, Xianyi Liu, Lijuan Zhang, Yun Li, Yu Liu, and Yuting Su. 2024. Multimodal Attentive Representation Learning for Micro-video Multi-label Classification. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 6 (2024), 1–23.
- [18] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.
- [19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [20] Jiayu Li, Hanyu Li, Zhiyu He, Weizhi Ma, Peijie Sun, Min Zhang, and Shaoping Ma. 2024. ReChorus2.0: A Modular and Task-Flexible Recommendation Library. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 454–464.
- [21] Nian Li, Xin Ban, Cheng Ling, Chen Gao, Lantao Hu, Peng Jiang, Kun Gai, Yong Li, and Qingmin Liao. 2024. Modeling User Fatigue for Sequential Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 996–1005.
- [22] Yuchen Li. 2024. Narrative and aesthetic features of micro-short plays on short video platforms: A case study of Douyin micro-short plays. *Advances in Economic Development and Management Research* 1, 2 (2024), 40–45.
- [23] Guanyu Lin, Chen Gao, Yu Zheng, Yinfeng Li, Jianxin Chang, Yanan Niu, Yang Song, Kun Gai, Zhiheng Li, Depeng Jin, et al. 2024. Inverse Learning with Extremely Sparse Feedback for Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 396–404.
- [24] Xiao Lin, Xiaokai Chen, Linfeng Song, Jingwei Liu, Biao Li, and Peng Jiang. 2023. Tree based progressive regression model for watch-time prediction in short-video recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4497–4506.
- [25] Huizhi Liu, Chen Li, and Lihua Tian. 2022. Multi-modal Graph Attention Network for Video Recommendation. In *2022 IEEE 5th International Conference on Computer and Communication Engineering Technology (CCET)*. IEEE, 94–99.
- [26] Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024. Multimodal recommender systems: A survey. *Comput. Surveys* 57, 2 (2024), 1–17.
- [27] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-video co-attention network for personalized micro-video recommendation. In *The world wide web conference*. 3020–3026.
- [28] Kelong Mao, Jieming Zhu, Liangcai Su, Guohao Cai, Yuru Li, and Zhenhua Dong. 2023. FinalMLP: An Enhanced Two-Stream MLP Model for CTR Prediction. *arXiv preprint arXiv:2304.00902* (2023).
- [29] Chang Meng, Ziqi Zhao, Wei Guo, Yingxue Zhang, Haolun Wu, Chen Gao, Dong Li, Xiu Li, and Ruiming Tang. 2023. Coarse-to-fine knowledge-enhanced multi-interest learning framework for multi-behavior recommendation. *ACM Transactions on Information Systems* 42, 1 (2023), 1–27.
- [30] Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. 2023. A content-driven micro-video recommendation dataset at scale. *arXiv preprint arXiv:2309.15379* (2023).
- [31] Emmanuel Opara, Theresa Mfon-Ette Adalikwu, and Caroline Aduke Tolorunleke. 2025. The Impact of Tiktok's Fast-Paced Content on Attention Span of Students. (2025).
- [32] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. FUM: fine-grained and fast user modeling for news recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1974–1978.
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [34] 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*. Springer, 261–278.
- [35] Lei Sang, Honghao Li, Yiwen Zhang, Yi Zhang, and Yun Yang. 2024. AdaGIN: Adaptive Graph Interaction Network for Click-Through Rate Prediction. *ACM Transactions on Information Systems* 43, 1 (2024), 1–31.
- [36] Yu Shang, Chen Gao, Jiansheng Chen, Depeng Jin, Meng Wang, and Yong Li. 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*. 433–442.
- [37] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 565–573.
- [38] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* 25 (2022), 5107–5116.
- [39] Yu Tian, Jianxin Chang, Yanan Niu, Yang Song, and Chenliang Li. 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*. 1632–1641.
- [40] Chenyang Wang, Yuanqing Yu, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. 2022. Towards representation alignment and uniformity in collaborative filtering. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 1816–1825.
- [41] Chenyang Wang, Min Zhang, Weizhi Ma, Yiqun Liu, and Shaoping Ma. 2020. Make it a chorus: knowledge- and time-aware item modeling for sequential recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 109–118.
- [42] Jing Wang, Aixin Sun, Hao Zhang, and Xiao Li. 2023. MS-DETR: Natural Language Video Localization with Sampling Moment-Moment Interaction. *arXiv preprint arXiv:2305.18969* (2023).
- [43] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*. 1785–1797.

- [44] Fanyue Wei, Biao Wang, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. 2022. Learning pixel-level distinctions for video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3073–3082.
- [45] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.
- [46] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. Multi-modal graph contrastive learning for micro-video recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1807–1811.
- [47] Ruohan Zhan, Changhua Pei, Qiang Su, Jianfeng Wen, Xueliang Wang, Guanyu Mu, Dong Zheng, Peng Jiang, and Kun Gai. 2022. Deconfounding duration bias in watch-time prediction for video recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4472–4481.
- [48] Weiqi Zhao, Dian Tang, Xin Chen, Dawei Lv, Daoli Ou, Biao Li, Peng Jiang, and Kun Gai. 2023. Disentangled causal embedding with contrastive learning for recommender system. In *Companion Proceedings of the ACM Web Conference 2023*. 406–410.
- [49] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.
- [50] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1059–1068.
- [51] Xin Zhou. 2023. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*. 1–2.
- [52] Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 935–943.
- [53] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.