

Exploiting Fine-Grained Skip Behaviors for Micro-Video Recommendation

Sanghyuck Lee, Sangkeun Park, Jaesung Lee*

Department of Artificial Intelligence, Chung-Ang University, Seoul, Republic of Korea
 {tkdgur658, psk492, curseor}@cau.ac.kr

Abstract

The growing trend of sharing short videos on social media platforms, where users capture and share moments from their daily lives, has led to an increase in research efforts focused on micro-video recommendations. However, conventional methods oversimplify the modeling of skip behavior, categorizing interactions solely as positive or negative based on whether skipping occurs. This study was motivated by the importance of the first few seconds of micro-videos, leading to a refinement of signals into three distinct categories: highly positive, less positive, and negative. Specifically, we classify skip interactions occurring within a short time as negatives, while those occurring after a delay are categorized as less positive. The proposed dual-level graph and hierarchical ranking loss are designed to effectively learn these fine-grained interactions. Our experiments demonstrated that the proposed method outperformed three conventional methods across eight evaluation measures on two public datasets.

Introduction

Micro-videos typically refer to self-generated video content that is usually less than three minutes long, covering a wide range of daily life aspects such as the latest news, funny clips, and sports highlights (Zhang, Wang, and Ariffin 2024). According to a report by Vidico, expenditures on micro-video advertisements are expected to reach approximately 100 billion dollars in 2024, while video content is projected to account for 82% of global internet traffic by 2025 (Chaves 2024). Furthermore, recent policies by major companies, such as the TikTok creator fund policy and YouTube creator partnership program, underscore the continued global expansion of investment in micro-videos (Perez 2024). Given this overwhelming abundance, access to micro-videos is primarily driven by recommendation algorithms rather than self-searching (Park 2023), and user satisfaction declines when platforms repeatedly display videos that do not align with their interests (Gu and Hu 2024). As a result, the need for highly sophisticated recommendation systems that can effectively analyze user preferences and identify potentially interesting micro-videos in a personalized manner has become even more critical (Lu et al. 2023).

*Corresponding author.

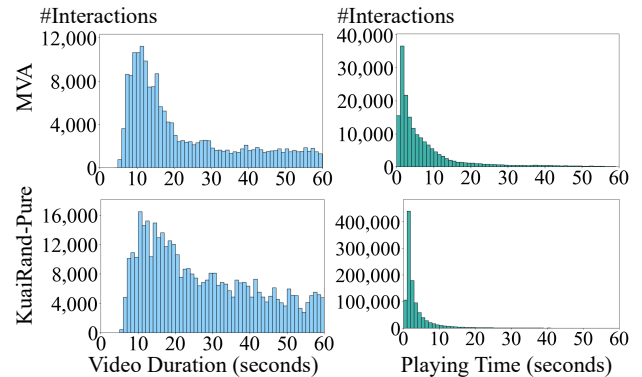


Figure 1: Distribution of video duration and playing time of the potent skipped interactions in two datasets, MVA (Shang et al. 2023) and KuaiRand-Pure (Gao et al. 2022b). The figures include only interactions where the playing time is shorter than the video duration; thus, playing time can be considered indicative of the timing of skip behaviors. Most skips occur within the first five seconds of the video, while the distribution of video durations remains relatively uniform. The conventional approach (Shang et al. 2023) based on playing time views incomplete viewing as negative, ignoring that users might form positive impressions early, causing slightly delayed skips. The histogram bin range has been truncated to 0-60 seconds for the sake of clarity.

Conventional approaches for micro-video recommendation are roughly divided into two strategies. One strategy involves utilizing the multi-modal information of micro-videos (Wei et al. 2019; Wang, Wu, and Hoashi 2019), and the other group aims to capture the different interests of users (Jiang et al. 2020; Tian et al. 2022). Despite their success, conventional methods have limited potential for further performance improvements due to their failure to effectively leverage the rich information that playing time conveys about both positive and negative interests, such as skip behavior (Gu and Hu 2024).

While browsing content or using the platform, users should watch for a few seconds before deciding whether to continue viewing the video or swipe down to move on to the next one (Gao et al. 2022a). Due to this skippable nature

of micro-videos, the significance of the first few seconds has been highlighted by multimedia stakeholders (Willis 2024; Wang 2021; Banerjee and Pal 2021; Jia et al. 2022). These initial moments in micro-videos should be carefully considered to encourage users to watch the video until the end (Kelemen 2023). Similarly, in Figure 1, the statistics of Micro-video-A (MVA) (Shang et al. 2023) and KuaiRand-Pure (Gao et al. 2022b) datasets show that most skips occurred within the first five seconds of interaction, suggesting a strong negative signal from users for the video content. In other words, a delayed skip may have different traits from this strong negative signal, which is our primary focus in this study.

Inspired by this motivation, we propose a dual-graph-based micro-video recommender, which contains a dual-level positive graph receiving help from the less positive interest separated from skip behavior interactions. Furthermore, negative interactions are integrated into the optimization process rather than being included in this graph construction, resulting in a hierarchical ranking loss. We summarize the contributions of this paper as follows:

- The proposed model, an adaptation of the conventional FRAME model that distinguishes between positive and negative interactions based on whether skipping occurs, refines the interaction types into three categories: highly positive, less positive, and negative. This improved approach demonstrates superior performance across eight evaluation measures on two datasets compared to the three conventional models.
- By considering the delayed skip as a less positive signal, the proposed dual-graphs using the dual-level positives demonstrate higher performance compared to both training with only the highly positive signal or training without distinguishing between the two levels.
- The quick skip is regarded as carrying a strong negative signal, which helps in improving training with conventional Bayesian Personalized Ranking (BPR) loss.

Related Work

Early recommendations for micro-videos relied on standard collaborative filtering systems, which modeled interactions based solely on user and video identifiers (IDs). These approaches have since evolved to better capture the dynamic nature of user interests over time. For example, THACIL (Chen et al. 2018) utilized a hierarchical attention mechanism to capture video characteristics and user preferences, while UHMAN (Liu et al. 2020) recommended hashtags by analyzing video keywords within user histories. Later models, such as ALPINE (Li et al. 2019) and MTIN (Jiang et al. 2020), were designed to track and model user preferences across various time frames. DMR (Lu et al. 2023) further introduced capabilities to capture both historical and predictive user interest trends.

User-item interactions have been shown to naturally form a bipartite graph, facilitating complex information extraction between nodes (Wang et al. 2019, 2020; Rendle et al. 2020). Given their inherent complexity, micro-videos necessitate

analyses that incorporate visual, acoustic, and textual characteristics. Recommendations based on graph convolution networks (GCNs) typically integrate user-item interactions with multi-modal data. MMGCN (Wei et al. 2019) exploits user-video bipartite graphs for each modality, enhancing user profiles through data aggregation from multi-hop neighboring nodes. DualGNN (Wang et al. 2021) introduced a preference learning module to fine-tune interest assessments across modalities, while ElimRec (Liu et al. 2022) applied causal inference to minimize biases associated with single-modality focus. Following models, such as HUIGN (Wei et al. 2021) and HGCL (Cai et al. 2022a), employed contrastive learning for hierarchical and heterogeneous understanding of user-video relationships, with A2BM2GL (Cai et al. 2022b) and LUDP (Lei et al. 2023) further refining these approaches by optimizing graph weights and user preference modeling. GRCN (Wei et al. 2020), CONDE (Liu et al. 2021), and HHFAN (Cai et al. 2021) have investigated techniques, such as graph refinement and subgraph construction to enhance computational efficiency and recommendation accuracy. In a different approach, FRAME (Shang et al. 2023) proposed a refined recommendation method that utilizes dual-graph construction with video clips labeled by user skip behaviors.

To address the limitations of supervised learning for interaction modeling, different strategies have been investigated. SLMRec (Tao et al. 2022) improved the representation of feature patterns using data augmentation and contrastive learning in different modalities. Similarly, MMGCL (Yi et al. 2022) applied data augmentation but introduced negative sampling techniques to enhance the learning of modality contribution and correlation. MMSSL (Wei et al. 2023a) employed adversarially trained transformed instances for cross-modal semantic similarity-based contrastive learning, enhancing model generalization and interaction capture. InvRL (Du et al. 2022) addressed biased correlations from diverse data usage by clustering user-video interactions into different environments, learning invariant representations in each to model user preferences with causal insight.

Conventional GCNs inherently treat all neighbors equally during information aggregation, which is a drawback, as they assign the same weight to all interactions. To address this issue, attention mechanisms-based GCNs have emerged. UVCAN (Liu et al. 2019) independently embedded user histories and video features, modeling their dynamic interactions through a co-attention mechanism. MGAT (Tao et al. 2020) built a bipartite graph based on interactions and aggregated weights through a modality-specific attention mechanism to discern user modal preferences. MMKGV (Liu, Li, and Tian 2022) also employed an attention mechanism, constructing a knowledge graph based on video similarities to weigh user interactions. LightGT (Wei et al. 2023b) utilized a transformer-based self-attention block to capture complex patterns and interactions between user-video nodes, effectively using layers.

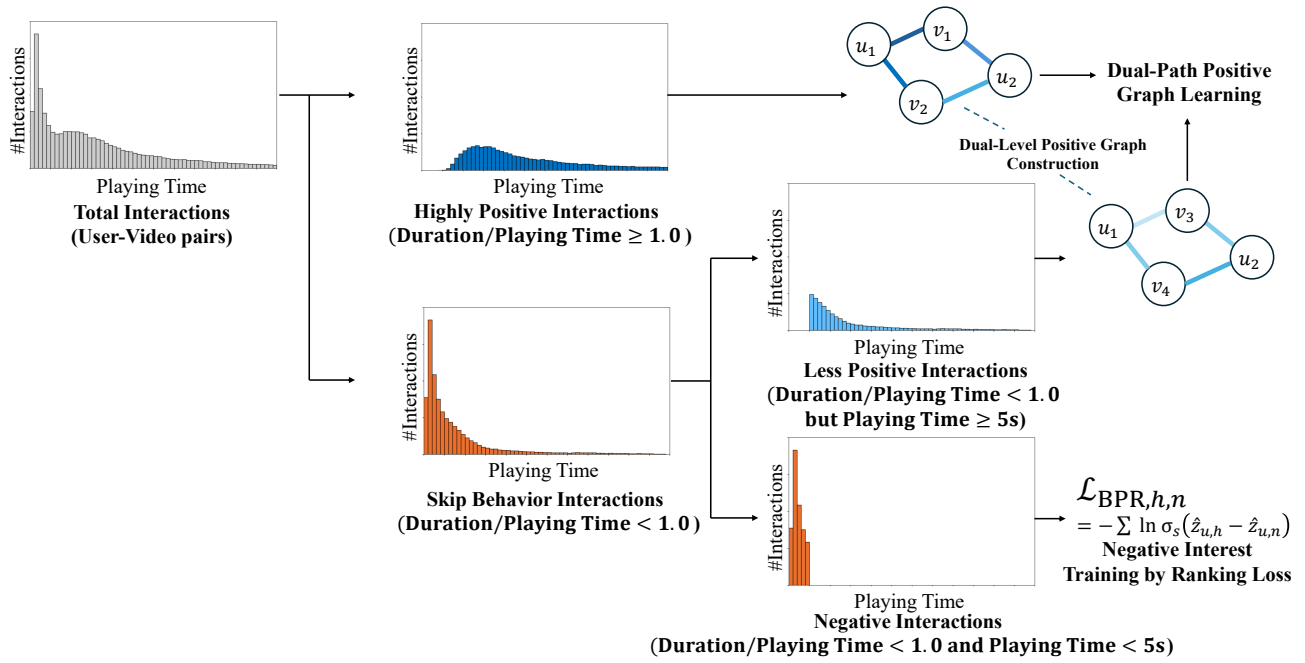


Figure 2: Dual-level positive graph construction and negative interest training by ranking loss. The total interactions are initially divided into Highly Positive Interactions and Skip Behavior Interactions based on Duration/Playing Time. Then, Skip Behavior Interactions are further divided into Less Positive Interactions and Negative Interactions, with Playing Time 5s as the threshold. Less Positive Interactions indicate a preference to continue watching the video beyond the initial 5 seconds, a period where skips are most frequent. Highly Positive Interactions and Less Positive Interactions each form individual adjacency graphs. These two adjacency graphs are utilized in dual-path positive graph learning. Negative Interactions help the model learn preference differences between interactions through a ranking loss.

Method

Problem Definition

Let $\mathcal{U} = \{u_1, u_2, \dots, u_{|\mathcal{U}|}\}$ and $\mathcal{V} = \{v_1, v_2, \dots, v_{|\mathcal{V}|}\}$ denote the user set and the video set, respectively. The total historical interactions between users and videos are formatted as a sequence of triplets, $\mathcal{I} = \langle (u_i, v_j, y_k)_{k=1}^{|\mathcal{I}|} \rangle$ where y_k represents the corresponding label indicating whether the video in the k -th interaction was 100% viewed or skipped by the user in the k -th interaction. The skip case is denoted as zero, and the fully viewed case is denoted as one. For example, $(u_1, v_3, 1)$ means that the user u_1 watched all of the video v_3 and $(u_2, v_5, 0)$ means that the user u_2 skip the video v_5 . The problem in this study can be formulated as follows.

Input: The total interactions \mathcal{I} ; The visual features X_v extracted from the image pixels of video frames by pre-trained image feature extraction model for all videos v in \mathcal{V} .

Output: A micro-video recommendation model that estimates the preference score of a user u given video v .

Proposed Method

Dual-Level Positive Graph Construction. To exploit less positive interest from users, we construct two separate user-item interaction graphs by leveraging the skip behaviors observed in the first few seconds of video playback. These graphs are designed to capture highly and less positive signals from user interactions, respectively.

Given the user u in the set of users as \mathcal{U} , the number of videos skipped within the first few seconds is denoted as $N_t^{(u)}$, while the remaining videos in the interactions are denoted as $N_h^{(u)}$. Duplicated videos caused by duplicate interaction are considered highly positive.

For the highly positive signal, we collect all videos from the interactions of the user u that were not skipped within the first few seconds to form the highly positive video set \mathcal{V}_u^h , which is defined as

$$\mathcal{V}_u^h = \{v_u^{h,1}, v_u^{h,2}, \dots, v_u^{h,|\mathcal{V}_u^h|}\}, \quad (1)$$

where each $v_u^{h,k}$ represents a video from the interaction history of user u that was not skipped in the first few seconds.

For the less positive signal, we focus on the videos during which the user u exhibited delayed skipping behavior after the first few seconds. This forms the less positive video set \mathcal{V}_u^l , defined as

$$\mathcal{V}_u^l = \{v_u^{l,1}, v_u^{l,2}, \dots, v_u^{l,|\mathcal{V}_u^l|}\}, \quad (2)$$

where each $v_u^{l,k}$ represents a video from the interaction history of user u that was skipped after the first few seconds.

Given the video sets with highly positive signal \mathcal{V}_u^h and less positive signal \mathcal{V}_u^l for each user, we construct two corresponding user-video interaction graphs. These graphs can be represented using the interaction matrices R^h and R^l for

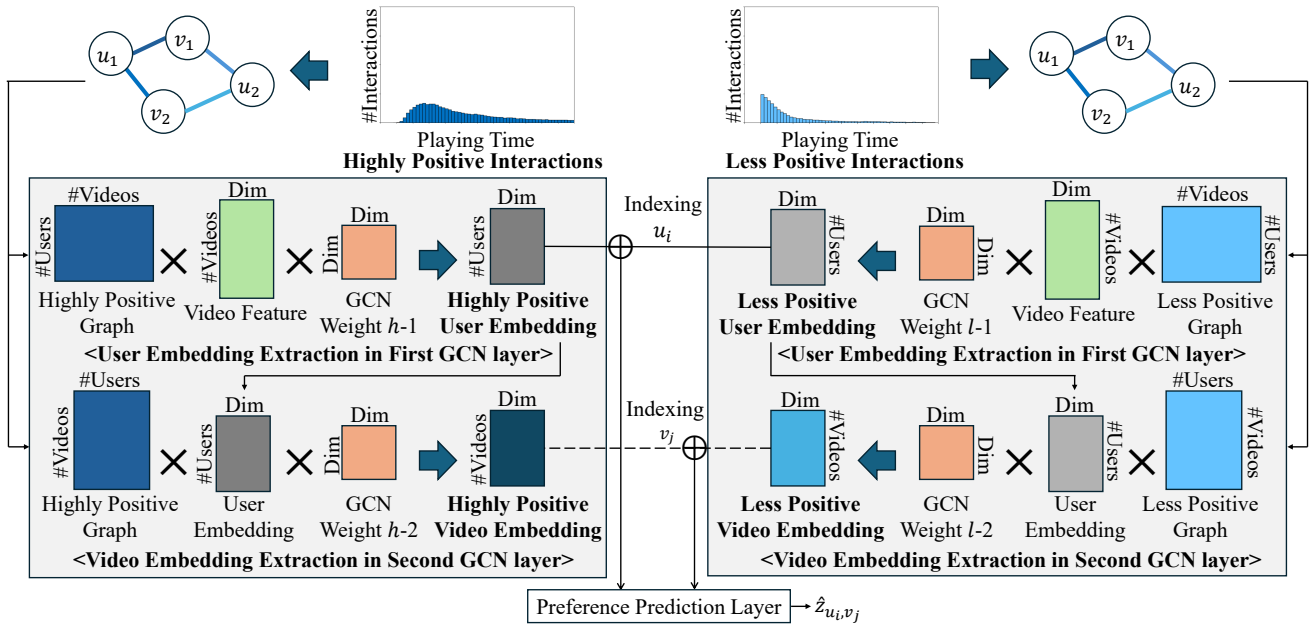


Figure 3: A schematic overview of the proposed dual-path positive graph learning. The video features are processed through distinct paths corresponding to the adjacency matrices from the highly and less positive graphs, reaching the preference prediction layer. The user embedding and video embedding generated from the two paths are then mean-pooled and concatenated. The fused features are passed through the prediction layer to output the preference score of the user u_i for the video v_j .

the highly and less positive relationships. For a set of users \mathcal{U} and videos \mathcal{V} , the interaction matrices R^h and R^l are defined as

$$R_{ij}^h = \begin{cases} 1 & \text{if } v_j \in \mathcal{V}_{u_i}^h; \\ 0 & \text{otherwise;} \end{cases} \quad (3)$$

whereas

$$R_{ij}^l = \begin{cases} 1 & \text{if } v_j \in \mathcal{V}_{u_i}^l; \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Using these interaction matrices, we construct the dual-side adjacency matrices for the user-video graphs as

$$A^h = \begin{pmatrix} 0 & R^h \\ (R^h)^\top & 0 \end{pmatrix}, \quad (5)$$

whereas

$$A^l = \begin{pmatrix} 0 & R^l \\ (R^l)^\top & 0 \end{pmatrix}. \quad (6)$$

These matrices are then normalized to construct the adjacency matrices \tilde{A}^h and \tilde{A}^l for the highly positive and less positive interactions, respectively. A symmetric normalization approach was used, where values were divided by the square root of the column and row degrees (He et al. 2020). The normalized adjacency matrices are given by

$$\tilde{A}^h = (D^h)^{-\frac{1}{2}} A^h (D^h)^{-\frac{1}{2}}, \quad (7)$$

whereas

$$\tilde{A}^l = (D^l)^{-\frac{1}{2}} A^l (D^l)^{-\frac{1}{2}}, \quad (8)$$

where D^h and D^l are diagonal matrices representing the degree of nodes in the highly and less positive graphs, respectively. Each entry D_{ii} in these matrices denotes the number

of non-zero entries in the i -th row of the corresponding adjacency matrix.

By constructing and normalizing these graphs, we effectively separate highly and less positive signals, where less positive signals in this study were treated as negative signals in conventional studies, based on user skip behaviors which will enable the recommendation model to better understand and predict user preferences for micro-video content.

Dual-Path Positive Graph Learning. Inspired by GCNs success in modeling higher-order interactions by aggregating information from different-hop neighbors, we extend this approach to user-video interactions. We derive user embeddings via the mechanism of embedding propagation in GCN models, where user embeddings are generated by aggregating the embeddings of their neighbors. For a user, neighbors are videos they have interacted with, while for a video, neighbors are users who interacted with it.

We compute two sets of embeddings for each user, corresponding to highly and less positive interactions, as

$$H_u^h = \sigma \left(\left(\tilde{R}^h \right)^\top H_v^{(0)} W_h^{(1)} \right), \quad (9)$$

whereas

$$H_u^l = \sigma \left(\left(\tilde{R}^l \right)^\top H_v^{(0)} W_l^{(1)} \right), \quad (10)$$

where $H_v^{(0)} \in \mathbb{R}^{|\mathcal{V}| \times d}$ denotes the initial video embedding matrix, which is derived from the visual features of the videos. The matrices \tilde{R}^h and \tilde{R}^l are the interaction matrices

defined in Equations (4) and (5), respectively, and $\sigma(\cdot)$ represents the nonlinear activation function. The matrices $W_h^{(1)}$ and $W_l^{(1)} \in \mathbb{R}^{d \times d}$ are trainable weight matrices. Thus, H_u^h and H_u^l are the user embeddings that capture highly and less positive interactions, respectively, and will be used for subsequent prediction tasks.

To capture higher-order relationships between videos, we perform a two-hop embedding propagation, which is formulated as

$$H_v^h = \sigma \left(\left(\tilde{R}^h \right)^\top H_u^h W_h^{(2)} \right), \quad (11)$$

whereas

$$H_v^l = \sigma \left(\left(\tilde{R}^l \right)^\top H_u^l W_l^{(2)} \right), \quad (12)$$

where $W_h^{(2)}$ and $W_l^{(2)} \in \mathbb{R}^{d \times d}$ are trainable weight matrices in the second GCN layer. Since both highly and less positive interactions represent multiple levels of user preference, we can combine these embeddings rather than treating them separately at this stage. Thus, we apply mean pooling to obtain the final user and video embedding representations, defined as

$$H_u = \text{Mean} \left(H_u^h, H_u^l \right), \quad (13)$$

whereas

$$H_v = \text{Mean} \left(H_v^h, H_v^l \right). \quad (14)$$

In summary, we obtain $H_u \in \mathbb{R}^{|\mathcal{U}| \times d}$ and $H_v \in \mathbb{R}^{|\mathcal{V}| \times d}$ as the embeddings of users and videos, respectively. These embeddings incorporate both highly and less positive interactions, allowing them to represent the users and videos collectively and effectively.

Preference Prediction Layer

After fusing dual-side interest embeddings of users and embeddings of videos from different GCN layers, we can now make the prediction. For a given user u_i and video v_j , we index the corresponding user embedding $\mathbf{h}_{u_i} \in \mathbb{R}^d$ and video embedding $\mathbf{h}_{v_j} \in \mathbb{R}^d$ from H_u and H_v , respectively. These embeddings are then concatenated and multiplied by a weight matrix, followed by a non-linear activation function, and finally multiplied by another weight matrix to output the final preference score

$$\hat{z}_{u_i, v_j} = W^{(2)} \cdot \sigma(W^{(1)} \cdot [\mathbf{h}_{u_i}, \mathbf{h}_{v_j}]), \quad (15)$$

where $W^{(1)} \in \mathbb{R}^{2d \times d}$ and $W^{(2)} \in \mathbb{R}^{d \times 1}$ are the weight matrices, and $[\cdot, \cdot]$ denotes the concatenation operation.

Optimization

BPR Loss with Highly/Less Positive and Negative Samples. In this study, we extend the traditional BPR loss by incorporating highly positive (s), less positive (w), and negative (n) samples. For each training interaction involving a user u and a video v , we construct triplets by sampling two additional items to get one each of highly/less positive and negative. This triplet-based sampling strategy ensures that the model learns both from strong preference interest and from comparisons between varying levels of user preference and non-preference.

Dataset	#Users	#Videos	#Interactions
Micro-video-A	12,739	58,291	342,694
KuaiRand-Pure (Random policy)	27,285	7,583	1,186,059

Table 1: Brief statistics of datasets employed in our study

The BPR loss is computed twice: once using the highly and less positive items and once using the highly positive and negative items. The final BPR loss is the average of these two, which helps in balancing the ability of the model to rank items within different levels of user preference. The BPR loss for the highly and less positive interactions is defined as

$$\mathcal{L}_{\text{BPR}, h, l} = - \sum_{(u, h, l) \in D_{h, l}} \ln \sigma_s(\hat{z}_{u, h} - \hat{z}_{u, l}), \quad (16)$$

where σ_s is the sigmoid function, $\hat{z}_{u, h}$ and $\hat{z}_{u, l}$ are the predicted scores for the highly and less positive items, respectively, for user u , and $D_{h, l}$ represents the set of all training triples (u, h, l) . This step ensures that the model can effectively rank items, even among those that the user has already shown some preference for, refining the precision of the recommendation system. Similarly, the BPR loss for the highly positive and negative items is defined as

$$\mathcal{L}_{\text{BPR}, h, n} = - \sum_{(u, h, n) \in D_{h, n}} \ln \sigma_s(\hat{z}_{u, h} - \hat{z}_{u, n}), \quad (17)$$

where $\hat{z}_{u, h}$ and $\hat{z}_{u, n}$ are the predicted scores for the highly positive and negative items, respectively, for user u , and $D_{h, n}$ represents the set of all training triples (u, h, n) . This step enhances the ability of the model to distinguish between items that are highly preferred and those that are not preferred at all, enhancing the discrimination power of the model. The overall BPR loss is then computed as the average of these two losses

$$\mathcal{L}_{\text{BPR}} = \frac{1}{2} (\mathcal{L}_{\text{BPR}, h, l} + \mathcal{L}_{\text{BPR}, h, n}). \quad (18)$$

Averaging these losses ensures that the model maintains a balanced perspective, optimizing both intra-preference ranking between highly and less positive items and inter-preference ranking between highly positive and negative items.

BCE Loss for Supervised Learning. To further enhance the supervision, we integrate binary cross-entropy (BCE) loss based on whether a video was skipped or not. BCE loss treats the problem as a binary classification task, where the label $y = 1$ indicates that a video was not skipped, and $y = 0$ indicates that the video was skipped. The BCE loss is defined as

$$\mathcal{L}_{\text{BCE}} = - [y \cdot \ln(\sigma_s(\hat{z})) + (1 - y) \cdot \ln(1 - \sigma_s(\hat{z}))], \quad (19)$$

where y is the true binary label, and \hat{z} is the logit, which is the output before applying the sigmoid function to obtain the predicted probability that the video was not skipped. This loss enables the model to rank videos while simultaneously learning to classify them correctly based on whether they were skipped, providing a supervised learning signal that complements the ranking provided by the BPR loss.

Model	Precision@3	Recall@3	MAP@3	NDCG@3	Precision@5	Recall@5	MAP@5	NDCG@5
Proposed	0.573*	0.623*	0.739*	0.790*	0.540*	0.882*	0.731*	0.812*
	(± 0.002)	(± 0.002)	(± 0.002)	(± 0.002)	(± 0.002)	(± 0.001)	(± 0.002)	(± 0.001)
BM3	0.546	0.591	0.701	0.758	0.532	0.869	0.699	0.787
	(± 0.003)	(± 0.003)	(± 0.005)	(± 0.004)	(± 0.002)	(± 0.001)	(± 0.003)	(± 0.003)
FRAME	0.538	0.581	0.697	0.753	0.528	0.863	0.694	0.784
	(± 0.004)	(± 0.004)	(± 0.004)	(± 0.004)	(± 0.002)	(± 0.001)	(± 0.003)	(± 0.003)
LightGT	0.506	0.563	0.660	0.720	0.505	0.856	0.664	0.760
	(± 0.003)	(± 0.003)	(± 0.004)	(± 0.004)	(± 0.001)	(± 0.001)	(± 0.003)	(± 0.003)

Table 2: The experimental results on the MVA dataset. The highest scores are marked in bold, with statistically significant paired t -test results ($p = 0.01$) indicated by an asterisk (*).

Model	Precision@3	Recall@3	MAP@3	NDCG@3	Precision@5	Recall@5	MAP@5	NDCG@5
Proposed	0.279*	0.632*	0.545*	0.591*	0.234*	0.760*	0.565*	0.637*
	(± 0.004)	(± 0.009)	(± 0.010)	(± 0.009)	(± 0.002)	(± 0.006)	(± 0.008)	(± 0.008)
BM3	0.214	0.497	0.386	0.433	0.198	0.646	0.417	0.495
	(± 0.014)	(± 0.029)	(± 0.037)	(± 0.037)	(± 0.009)	(± 0.027)	(± 0.035)	(± 0.035)
FRAME	0.263	0.606	0.501	0.551	0.227	0.744	0.526	0.604
	(± 0.006)	(± 0.013)	(± 0.015)	(± 0.015)	(± 0.003)	(± 0.009)	(± 0.014)	(± 0.013)
LightGT	0.169	0.411	0.291	0.335	0.170	0.564	0.326	0.402
	(± 0.002)	(± 0.006)	(± 0.003)	(± 0.004)	(± 0.001)	(± 0.005)	(± 0.003)	(± 0.003)

Table 3: The experimental results on the KuaiRand-Pure dataset. The highest scores are marked in bold, with statistically significant paired t -test results ($p = 0.01$) indicated by an asterisk (*).

Combined Loss. The final combined loss function integrates the averaged BPR loss with the BCE loss, enabling the model to learn from both ranking and classification perspectives. The combined loss is given by

$$\mathcal{L}_{\text{combined}} = \lambda \mathcal{L}_{\text{BPR}} + (1 - \lambda) \mathcal{L}_{\text{BCE}}, \quad (20)$$

where λ is a hyperparameter that balances the contributions of the BPR loss and the BCE loss.

By combining these losses, the model benefits from the strengths of both approaches: the BPR loss ensures effective ranking of items according to nuanced user preferences, while the BCE loss supervises the model in accurately classifying videos based on user skip behavior. This comprehensive learning process results in a more refined and accurate recommendation system capable of both ranking items and predicting user engagement.

Experiments

Experimental Settings

Datasets. Table 1 shows brief statistics of two datasets. The MVA dataset (Shang et al. 2023) was collected from a mobile app platform. MVA includes interaction records containing user IDs, video IDs, playing time, video duration, interaction timestamps, and multi-level user behaviors such as likes, follows, and forwards. This dataset provides a rich context for analyzing user behavior by leveraging playing time and video duration to calculate the skip time for each user. For visual features, a pre-trained convolutional network is first used to extract features from each frame of all videos (Shang et al. 2023). Then, K frames are sampled from each video, and features are extracted for each frame. These features are averaged to form

a 128-dimensional vector. The MVA dataset consists of 12,739 users, 58,291 videos, and 342,694 interactions. The KuaiRand-Pure dataset (Gao et al. 2022b), collected from the Kuaishou app, one of the largest video-sharing platforms in China, provides an unbiased sequential recommendation dataset. This dataset distinguishes itself by intervening in the recommendation policies of the platform through random insertion of selected videos over a two-week period, allowing for the collection of genuine user feedback without their awareness. The dataset captures 12 types of feedback signals, including clicks, favorites, and view time. The KuaiRand-Pure dataset includes 27,285 users, 7,583 videos, and 1,186,059 interactions, depending on the recommendation policy applied. Since the KuaiRand-Pure dataset does not provide video features, we set up learnable parameters for node embeddings.

Baselines. We employed three state-of-the-art baselines in our experiments. Specifically, we used two multi-modal micro-video recommendation models, FRAME (Shang et al. 2023) and LightGT (Wei et al. 2023b), and one multi-media recommendation model, BM3 (Zhou et al. 2023). FRAME constructs a positive-negative graph for clip-level learning and models a dual-side GCN layer. LightGT inherits from the LightGCN model and Transformer, developing a modal-specific embedding and a layer-wise position encoder. BM3 bootstraps latent contrastive views in user-item representations, utilizing dropout augmentation.

Implementation Details. All models were implemented using PyTorch 2.3. The entire set of interactions is randomly sampled at a 6:2:2 ratio for each user to generate the training, validation, and test sets. Each model utilizes the train-

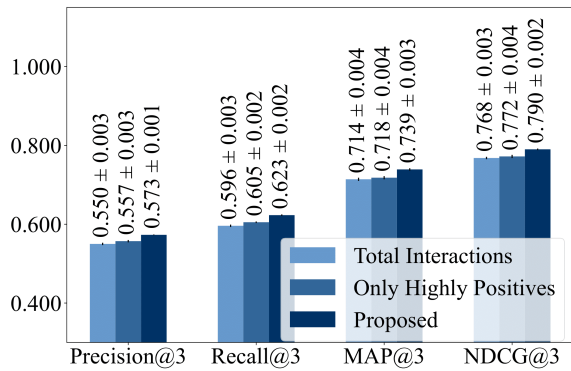


Figure 4: Comparison results between total interaction, highly positive only, and proposed dual-level graph.

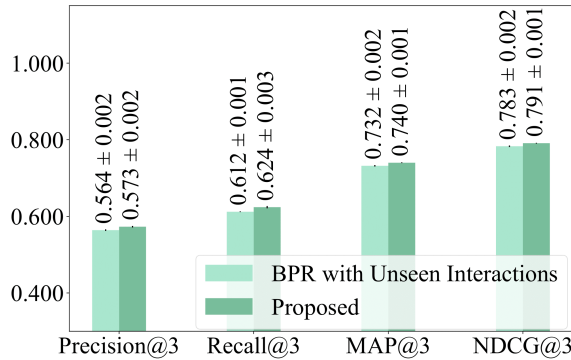


Figure 5: Comparison results between BPR loss with unseen interactions and proposed BRP loss.

ing set to construct the adjacency matrix and is trained for up to 30 epochs. Early stopping is applied if the recall@3 on the validation set does not improve for five consecutive epochs. The evaluation measures include Top- k recall, precision, mean average precision (MAP), and normalized discounted cumulative gain (NDCG) at $k = \{3, 5\}$. Preference prediction is conducted on the items associated with users in the test interactions.

To ensure robust results in a statistical way, data splitting, training, and testing were repeated ten times in the MVA dataset and seven times in the KuaiRand-Pure dataset. For performance comparison, statistical significance was assessed using a paired t -test implemented via the SciPy open-source library. Each model was trained with a batch size of 1024 using the AdamW optimizer (Loshchilov and Hutter 2019) with a momentum of 0.9 and a weight decay of $1e-4$. The learning rate starts at $1e-3$ and decays to $1e-6$ following a cosine annealing schedule. Each model was trained based on the loss function suggested in the original papers. The feature dimension d is set to 128, and λ is set to 0.5.

Experimental Results

Comparison Results. As shown in Tables 2 and 3, the two experimental results demonstrate that the proposed method outperforms the comparison models across two different

datasets. Table 2 presents results on the MVA dataset, the proposed method achieved the highest scores in all evaluation measures. Notably, the proposed method significantly outperformed other models with statistical significance at the $p=0.01$ level across all measures. Table 3 shows the results on the KuaiRand-Pure dataset, where the proposed method again achieved the best performance in all evaluation measures. In this dataset, the proposed method also significantly outperformed the other models with a $p=0.01$ level of statistical significance. Both experiments consistently demonstrate the superior performance of the proposed method compared to the conventional models. This consistent superiority across different datasets highlights the efficacy of the proposed method.

Ablation Study. Figures 4 and 5 illustrate the experimental results of the ablation study on the proposed model. As shown in Figure 4, we examined the performance of different graph construction methods. In the field of micro-video recommendation, the model can be trained using all interactions as positive signals, especially when there is a lack of explicit feedback or playing time data. Alternatively, playing time can be used to filter out data where skipping occurs, allowing the model to learn high-confidence user preferences. The proposed dual-level positive graph construction method outperformed both of these approaches. Using total interaction data without distinguishing is overly naive and can hinder the learning process. Furthermore, the relatively superior performance of the proposed model compared to the only highly approach suggests that less positive interactions can provide valuable information for interaction modeling.

As seen in Figure 5, we compared the widely used negative sampling for unseen interactions with the proposed BPR loss. The proposed BPR loss demonstrated superior performance, indicating that negative interactions that mean quick skip have a clearly defined relative ranking compared to highly positive interactions that were fully viewed. Furthermore, the dual BPR loss, which includes the less positive interactions, constructs a hierarchical ranking, suggesting that it warrants further investigation in future research.

Conclusion

This study proposes a dual-graph-based micro-video recommender system that effectively utilizes the granular details of user interactions, particularly by distinguishing based on skip behaviors between fully-viewed interactions, delayed skips, and quick skips. The experimental results demonstrate that our approach outperforms three conventional methods across eight evaluation measures on two public micro-video datasets.

In future research, we aim to develop methods that are not dependent on specific thresholds, such as 5 seconds, enabling their application to a wide range of datasets. For instance, in the KuaiRec dataset (Gao et al. 2022a), many skips occur even after 10 seconds. This indicates that user behavior may vary depending on the platform. In addition, advanced designs such as attention mechanisms (Tao et al. 2020; Liu, Li, and Tian 2022) can be incorporated into sub-modules to enhance the efficacy of the proposed model.

Acknowledgments

This research was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (2021-0-01341, Artificial Intelligence Graduate School Program (Chung-Ang University)), and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00459387). We are grateful to the reviewers for their insightful suggestions during the rebuttal process, which enhanced the quality of this paper and provided guidance for future studies.

References

- Banerjee, S.; and Pal, A. 2021. Skipping Skippable Ads on YouTube: How, When, Why and Why Not? In *Proceedings of 2021 15th International Conference on Ubiquitous Information Management and Communication*, 1–5.
- Cai, D.; Qian, S.; Fang, Q.; Hu, J.; Ding, W.; and Xu, C. 2022a. Heterogeneous Graph Contrastive Learning Network for Personalized Micro-Video Recommendation. *IEEE Transactions on Multimedia*, 25: 2761–2773.
- Cai, D.; Qian, S.; Fang, Q.; Hu, J.; and Xu, C. 2022b. Adaptive Anti-Bottleneck Multi-Modal Graph Learning Network for Personalized Micro-Video Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 581–590.
- Cai, D.; Qian, S.; Fang, Q.; and Xu, C. 2021. Heterogeneous Hierarchical Feature Aggregation Network for Personalized Micro-Video Recommendation. *IEEE Transactions on Multimedia*, 24: 805–818.
- Chaves. 2024. 20+ Interesting Short Form Video Statistics & Trends (2024). <https://vidico.com/news/short-form-video-statistics>. Accessed: 2024-08-14.
- Chen, X.; Liu, D.; Zha, Z.-J.; Zhou, W.; Xiong, Z.; and Li, Y. 2018. Temporal Hierarchical Attention at Category- And Item-Level for Micro-Video Click-Through Prediction. In *Proceedings of the 26th ACM International Conference on Multimedia*, 1146–1153.
- Du, X.; Wu, Z.; Feng, F.; He, X.; and Tang, J. 2022. Invariant Representation Learning for Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 619–628.
- Gao, C.; Li, S.; Lei, W.; Chen, J.; Li, B.; Jiang, P.; He, X.; Mao, J.; and Chua, T.-S. 2022a. KuaiRec: A Fully-Observed Dataset and Insights for Evaluating Recommender Systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 540–550.
- Gao, C.; Li, S.; Zhang, Y.; Chen, J.; Li, B.; Lei, W.; Jiang, P.; and He, X. 2022b. KuaiRand: An Unbiased Sequential Recommendation Dataset With Randomly Exposed Videos. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3953–3957.
- Gu, P.; and Hu, H. 2024. A Holistic View on Positive and Negative Implicit Feedback for Micro-Video Recommendation. *Knowledge-Based Systems*, 284: 111299.
- He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 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.
- Jia, W.; Zhou, R.; Chen, N.; and Shi, Y. 2022. Examining the Usability of a Short-Video App Interface Through an Eye-Tracking Experiment. In Soares, M. M.; Rosenzweig, E.; and Marcus, A., eds., *Design, User Experience, and Usability: UX Research, Design, and Assessment*, 414–427. Cham.
- Jiang, H.; Wang, W.; Wei, Y.; Gao, Z.; Wang, Y.; and Nie, L. 2020. What Aspect Do You Like: Multi-Scale Time-Aware User Interest Modeling for Micro-Video Recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3487–3495.
- Kelemen. 2023. Reach More People by Optimizing the First 5 Seconds of Your Video. <https://www.wintersummer.ca/blog/the-first-five-seconds-of-your-video-are-most-important>. Accessed: 2024-08-14.
- Lei, F.; Cao, Z.; Yang, Y.; Ding, Y.; and Zhang, C. 2023. Learning the User’s Deeper Preferences for Multi-Modal Recommendation Systems. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3s): 1–18.
- Li, Y.; Liu, M.; Yin, J.; Cui, C.; Xu, X.-S.; and Nie, L. 2019. Routing Micro-Videos via a Temporal Graph-Guided Recommendation System. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1464–1472.
- Liu, H.; Li, C.; and Tian, L. 2022. Multi-Modal Graph Attention Network for Video Recommendation. In *2022 IEEE International Conference on Computer and Communication Engineering Technology*, 94–99.
- Liu, S.; Chen, Z.; Liu, H.; and Hu, X. 2019. User-Video Co-Attention Network for Personalized Micro-Video Recommendation. In *Proceedings of the World Wide Web Conference*, 3020–3026.
- Liu, S.; Xie, J.; Zou, C.; and Chen, Z. 2020. User Conditional Hashtag Recommendation for Micro-Videos. In *Proceedings of the 2020 IEEE International Conference on Multimedia and Expo*, 1–6.
- Liu, X.; Tao, Z.; Shao, J.; Yang, L.; and Huang, X. 2022. ElimRec: Eliminating Single-Modal Bias in Multimedia Recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 687–695.
- Liu, Y.; Liu, Q.; Tian, Y.; Wang, C.; Niu, Y.; Song, Y.; and Li, C. 2021. Concept-Aware Denoising Graph Neural Network for Micro-Video Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1099–1108.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representation*, 1–18.
- Lu, Y.; Huang, Y.; Zhang, S.; Han, W.; Chen, H.; Fan, W.; Lai, J.; Zhao, Z.; and Wu, F. 2023. Multi-Trends Enhanced Dynamic Micro-Video Recommendation. In *Proceedings of the International Conference on Artificial Intelligence*, 430–441.

- Park. 2023. Social Media/Search Portal Trend Report. <https://blog.opensurvey.co.kr/trendreport/socialmedia-2023>. Accessed: 2024-08-14.
- Perez. 2024. YouTube Says Over 25 <https://techcrunch.com/2024/03/28/youtube-says-over-25-of-its-creator-partners-now-monetize-via-shorts>. Accessed: 2024-08-14.
- Rendle, S.; Krichene, W.; Zhang, L.; and Anderson, J. 2020. Neural Collaborative Filtering vs. Matrix Factorization Revisited. In *Proceedings of the 14th ACM Conference on Recommender Systems*, 240–248.
- Shang, Y.; Gao, C.; Chen, J.; Jin, D.; Wang, M.; and Li, Y. 2023. Learning Fine-Grained User Interests for Micro-Video Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 433–442.
- Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T.-S. 2022. Self-Supervised Learning for Multimedia Recommendation. *IEEE Transactions on Multimedia*, 25: 5107–5116.
- Tao, Z.; Wei, Y.; Wang, X.; He, X.; Huang, X.; and Chua, T.-S. 2020. Mgat: Multimodal Graph Attention Network for Recommendation. *Information Processing & Management*, 57(5): 102277.
- Tian, Y.; Chang, J.; Niu, Y.; Song, Y.; and Li, C. 2022. When Multi-Level Meets Multi-Interest: A Multi-Grained Neural Model for Sequential Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1632–1641.
- Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; and Nie, L. 2021. Dualgcn: Dual Graph Neural Network for Multimedia Recommendation. *IEEE Transactions on Multimedia*, 25: 1074–1084.
- Wang, X.; He, X.; Wang, M.; Feng, F.; and Chua, T.-S. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 165–174.
- Wang, X.; Jin, H.; Zhang, A.; He, X.; Xu, T.; and Chua, T.-S. 2020. Disentangled Graph Collaborative Filtering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1001–1010.
- Wang, Y. 2021. Content Characteristics and Limitations of Original Short Video Based on Depth Data. *Journal of Physics: Conference Series*, 1881(4): 042070.
- Wang, Y.; Wu, J.; and Hoashi, K. 2019. Multi-Attention Fusion Network for Video-Based Emotion Recognition. In *2019 International Conference on Multimodal Interaction*, 595–601.
- Wei, W.; Huang, C.; Xia, L.; and Zhang, C. 2023a. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*, 790–800.
- Wei, Y.; Liu, W.; Liu, F.; Wang, X.; Nie, L.; and Chua, T.-S. 2023b. LightGT: A Light Graph Transformer for Multimedia Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1508–1517.
- Wei, Y.; Wang, X.; He, X.; Nie, L.; Rui, Y.; and Chua, T.-S. 2021. Hierarchical User Intent Graph Network for Multimedia Recommendation. *IEEE Transactions on Multimedia*, 24: 2701–2712.
- Wei, Y.; Wang, X.; Nie, L.; He, X.; and Chua, T.-S. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation With Implicit Feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3541–3549.
- Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019. MMGCN: Multi-Modal Graph Convolution Network for Personalized Recommendation of Micro-Video. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1437–1445.
- Willis, M. 2024. *Short Video Adverts: A Modern and Virtual Form of Advertising*, 107–131. Cham: Springer International Publishing.
- Yi, Z.; Wang, X.; Ounis, I.; and Macdonald, C. 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.
- Zhang, Q.; Wang, Y.; and Ariffin, S. K. 2024. Keep Scrolling: An Investigation of Short Video Users’ Continuous Watching Behavior. *Information & Management*, 61(6): 104014.
- Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023. Bootstrap Latent Representations for Multi-Modal Recommendation. In *Proceedings of the ACM Web Conference 2023*, 845–854.