Master's Thesis Presentation for the 144th Dissertation Evaluation

A Study on Dual-Path Graph Convolution Network Using Skip Graph Separation for Micro-video Recommendation

마이크로 비디오 추천을 위한 스킵 그래프 분리를 이용한 이중 경로 그래프 합성곱 신경망 연구

Chaewoon Kim

rlacodns7@gmail.com

2025.09.11.

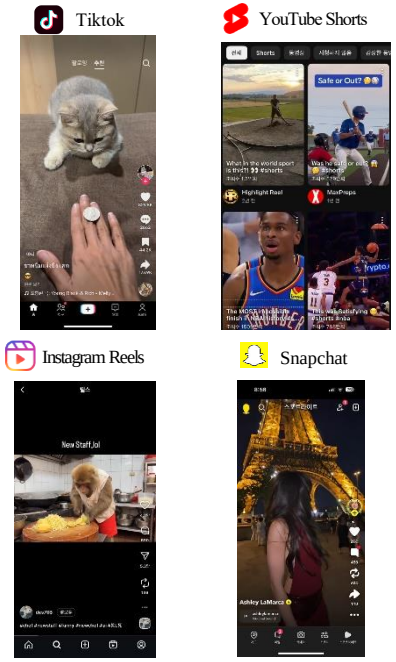
Contents

Introduction

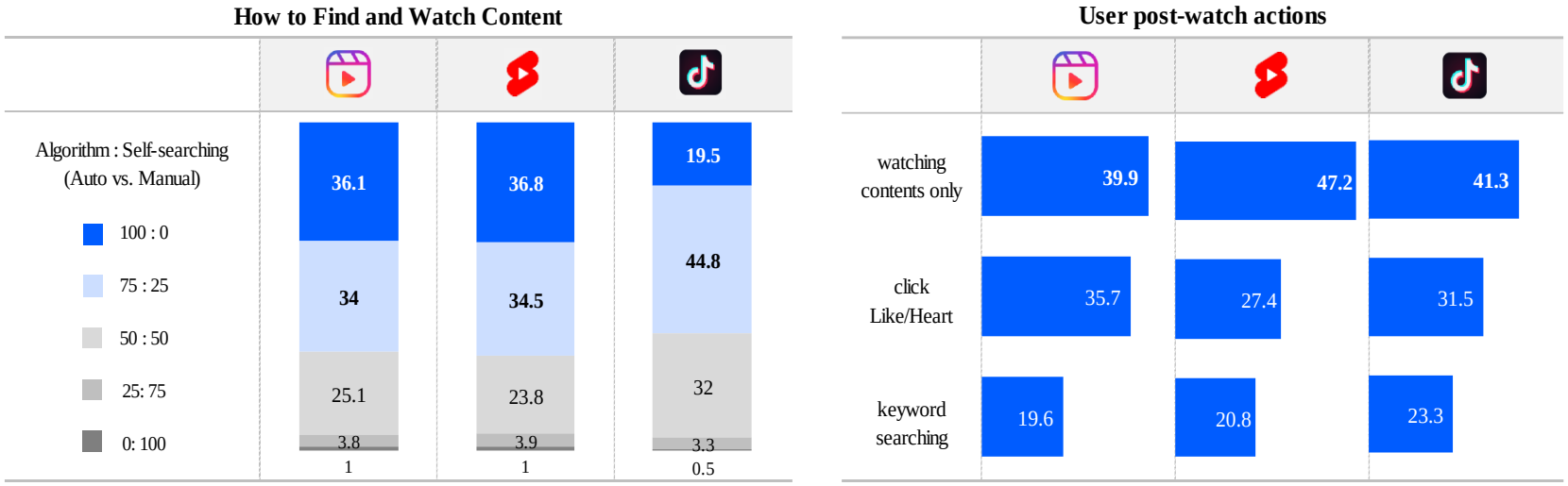
[1] S. Liu et al., "User Conditional Hashtag Recommendation for Micro-Videos," *2020 IEEE Int. Conf. Multimed.* Expo, London, UK, Jul. 2020, pp. 1-6.
[2] A. Chaudhary et al., "“Are you still watching?”: exploring unintended user behaviors and dark patterns on video streaming platforms.“ in *Proc. ACM Design. Interac. Sys. Conf.*, Australia, 2022, pp. 776-791.

Micro-video Recommendation System (MvRS)

- Micro-videos, also known as short-form content, are brief video pieces created by individuals that cover diverse aspects such as latest news, daily life clips, and sports highlights, rapidly increasing in recent years due to easy shareability on digital platforms [1].
- With improved content availability, micro-video exploration is driven by automatic playback recommendation, displaying content that fails to match user interests without skipping, requiring a sophisticated micro-video recommendation system (MvRS) [2].



Examples of Major Platforms for Micro-video



Statistics of Micro-video Usage in Major Platforms

Introduction

[3] [3] Z. Fu et al. "Deep Learning Models for Serendipity Recommendations: A Survey and New Perspectives." *ACM Comput. Surv.*, vol. 56, no. 19, Aug. 2023, pp. 1-26.
[4] M. Wilhelm et al., "Practical Diversified Recommendations on YouTube with Determinantal Point Processes," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manag.*, Torino, Italy, 2018, pp. 2165-2173.
[5] D. Klug et al., "Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm," in *Proc. 13th ACM Web Sci. Conf.*, United Kingdom, 2021, pp. 84-92.
[6] M. Boeker et al., "An Empirical Investigation of Personalization Factors on TikTok," in *Proc. ACM Web Conf.*, Lyon, France, 2022, pp. 2298-2309.

Objective and Standard Procedure

- MvRS aim to identify and recommend tailored video content that aligns with individual user interests, thereby enhancing user engagement and satisfaction by overcoming information overload and facilitating novel content discovery [3].
- Conventional micro-video recommendation models process historical user-video interactions as input, learn embeddings to capture the latent features of users and videos, and employ a scoring function to output a ranked list of recommendations [4, 5, 6].

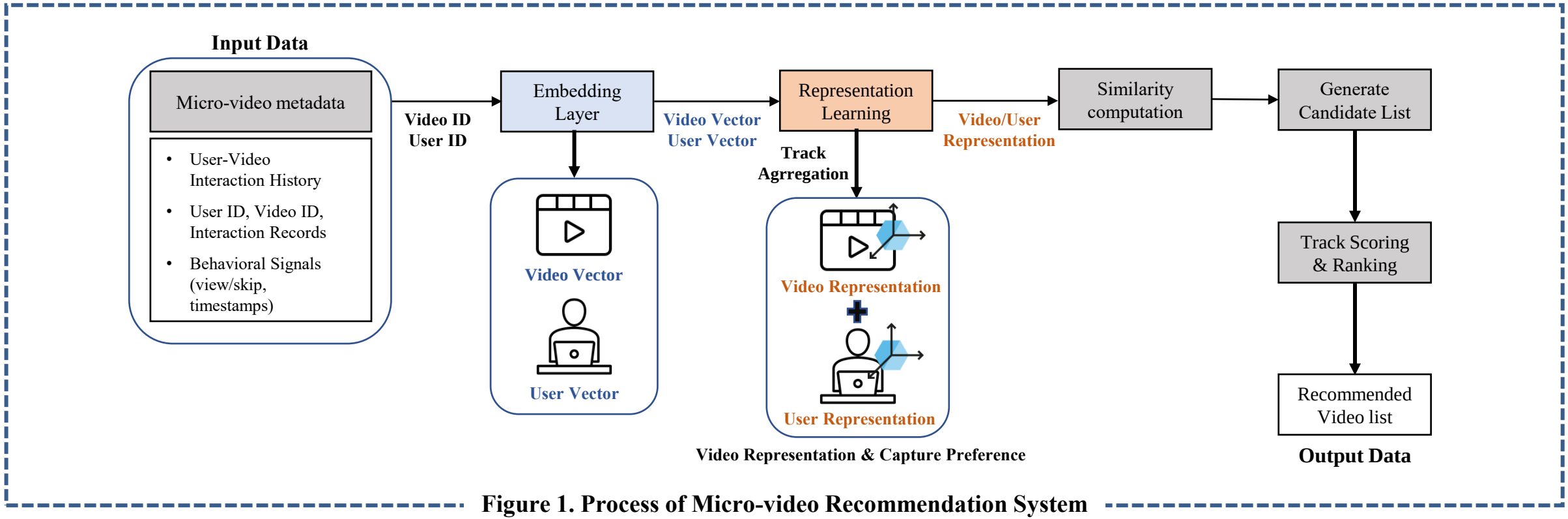


Figure 1. Process of Micro-video Recommendation System

Introduction

Ideal Goal and Challenge

- The primary goal of an MvRS is to model the intricate preferences of users by learning temporal engagement patterns, such as viewing duration and skip behavior, from their rich implicit signals, which enables the construction of more effective embeddings.
- Skip behavior only reveals whether users skipped or not, but users may skip immediately due to disinterest or after viewing engaging content, creating challenges in distinguishing and capture these different reasons.

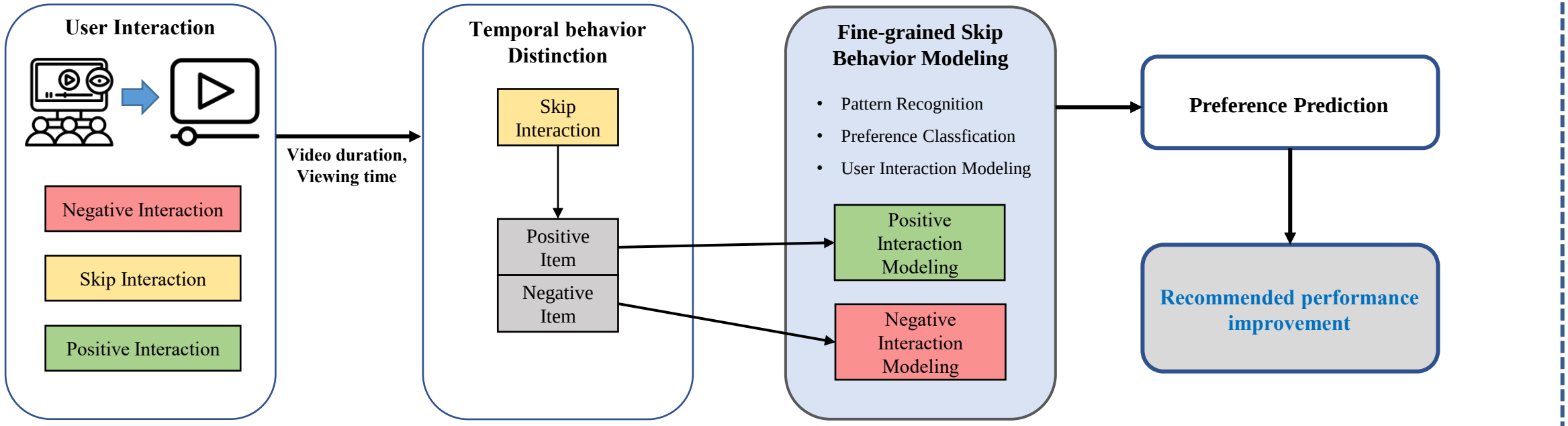


Figure 2. Performance Improvement with Preference Level distinction

[7] Y. Shang et al., "Learning Fine-grained User Interests for Micro-video Recommendation," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Taipei, Taiwan, 2023, pp. 433-442.

Related Work

FRAME(2023) [7]

- Features extracted from video clips are segregated into two sets based on complete viewing and skipping, forming positive and negative interaction graphs, and constructing clip-level learning through dual-side GCN layers for recommendation modeling.
- The approach to modeling negative interactions in FRAME is oversimplified, as it only treats the abandonment point as a negative signal and thereby overlooks the positive preference inherent in the content a user watched before skipping.

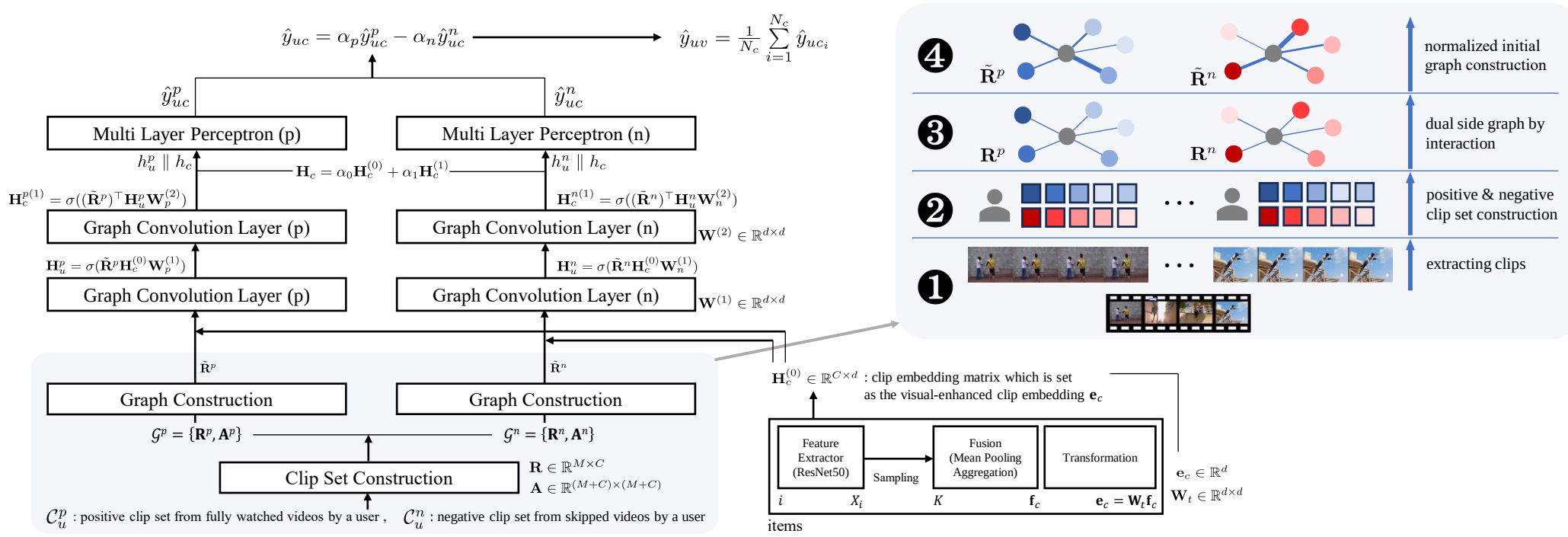


Figure 3. Illustration of the FRAME model

[8] Y. Wei et al., "LightGT: A Light Graph Transformer for Multimedia Recommendation," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Taipei, Taiwan, 2023, pp. 1508-1517.

Related Work

LightGT(2023) [8]

- Creates bipartite graphs from video features and users with viewing histories, capturing patterns and interactions between user-video nodes to facilitate structural information learning through common neighbors and multi-hop relationships.
- Fails to consider temporal engagement signals such as viewing duration or skip timing, treating all observed viewing behaviors equally as positive signals, which can result in negative interactions being misclassified as positive during preference learning.

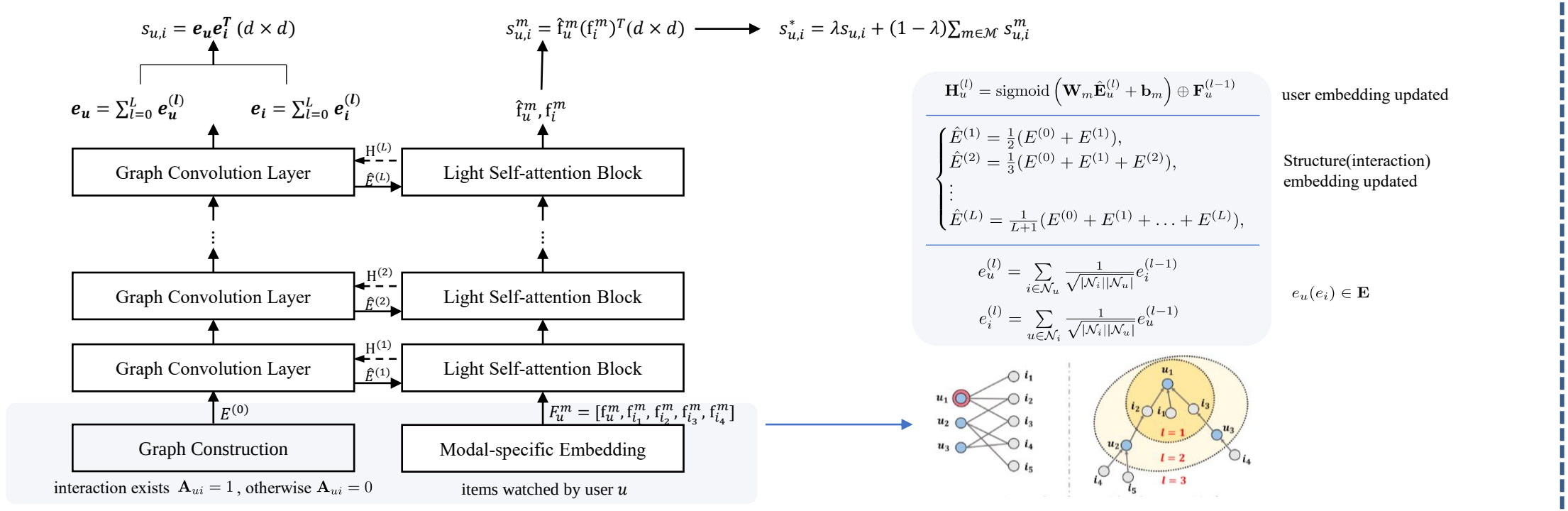


Figure 4. Overall Workflow of the LightGT Model

[9] Zhou. Xin et al. "Bootstrap latent representations for multi-modal recommendation." In *Proceedings of the ACM web conference 2023*, pp. 845-854.

Related Work

BM3(2023) [9]

- Generates contrastive views through simple dropout augmentation instead of graph augmentation, and optimizes graph reconstruction loss, inter-modality feature alignment loss, and intra-modality feature masking loss through multi-modal contrastive loss.
- Fails to consider temporal engagement signals and treats all observed user-item interactions equally as positive signals, thereby limiting the ability to distinguish fine-grained preferences during user preference representation learning.

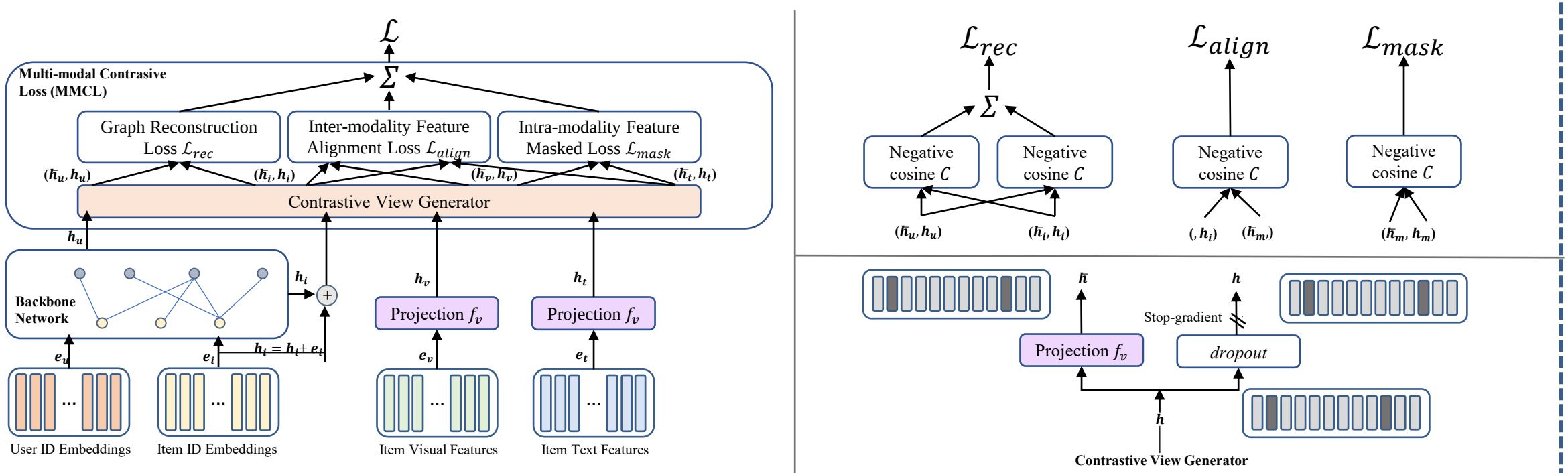


Figure 5. The structure overview of the BM3

[10] Yang. L et al. "Dgrec: Graph neural network for recommendation with diversified embedding generation." In Proceedings of the sixteenth ACM international conference on web search and data mining, 2023, pp. 661-669.

Related Work

DGRec(2023) [10]

- Integrates three modules that include submodular neighbor selection for diverse neighbor aggregation, layer attention to weight GNN layers, and loss reweighting for long-tail categories, to balance recommendation diversity and accuracy.
- Operating on binary interaction graphs with submodular selection focusing on embedding-space diversity rather than temporal engagement signals, thus failing to distinguish between immediate skip and delayed skip interactions during preference learning.

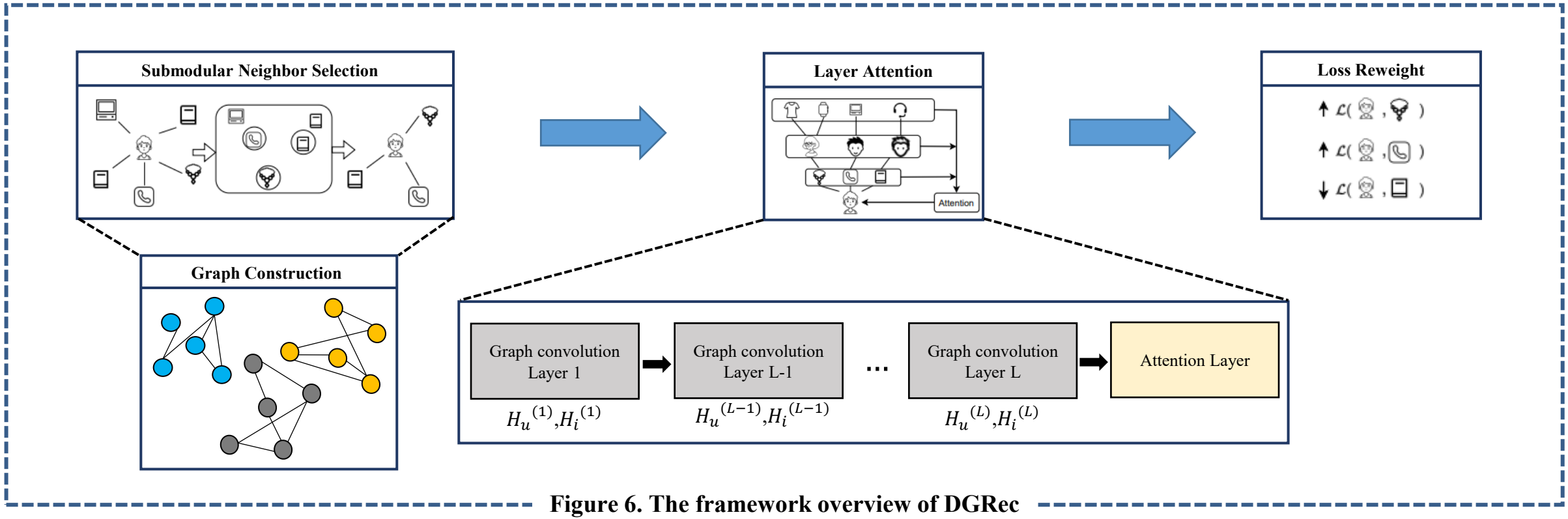


Figure 6. The framework overview of DGRec

Related Work

Common Drawback and Hypothesis

- Existing models either fail to utilize temporal engagement signals or rely on binary classification (viewed/skipped) for processing skip behavior, causing all intermediate-level preference signals to be regarded as negative and limiting accurate embedding learning.
- To avoid misclassification and loss of preference information, we conduct separate graph modeling for items involved in intermediate-level interactions. Subsequently, mean pooling is applied to calculate the final score, providing fine-grained preference learning.

Model	Author	Publication	Summary	Common Drawback
FRAME	Y. Shang et al. (2023)	46th International ACM SIGIR Conference on Research and Development in Information Retrieval	• Constructs dual-side GCN layers with positive and negative interaction graphs formed by segregating video clips based on complete viewing versus skipping behaviors for clip-level learning.	**Existing models either fail to utilize temporal engagement signals or rely on binary classification (viewed/skipped) for processing skip behavior, causing all intermediate-level preference signals to be regarded as negative and limiting accurate embedding learning.**
LightGT	Y. Wei et al. (2023)	46th International ACM SIGIR Conference on Research and Development in Information Retrieval	• Inherits LightGCN and Transformer architectures to develop modal-specific embeddings with layer-wise position encoders, creating bipartite graphs from video features and user viewing histories.	
BM3	Zhou, Xin, et al. (2023)	ACM Web Conference 2023 (WWW '23)	• Bootstraps latent contrastive views through dropout augmentation while jointly optimizing graph reconstruction, inter-modality alignment, and intra-modality masking losses.	
DGRec	Yang, L, et al. (2023)	16th ACM international conference on web search and data mining	• Integrates submodular neighbor selection, layer attention, and category-based loss reweighting to balance recommendation accuracy and diversity.	

Proposed Method

Notation and Definition

- $\mathcal{U}, \mathcal{I}, \mathcal{R}$ are collections of data, where the embeddings of user $\mathcal{U}$ and item $\mathcal{I}$ are denoted as $\mathbf{e}$, and the modality-specific features of $\mathcal{I}$ are represented by $\mathbf{f}$. The relationships formed between u and i, included in $\mathcal{R}$, contain preference labels y with watching time ratio.
- To utilize micro-video aspects, item i is divided into clips c, and clip set C_u are constructed based on interactions $\mathcal{R}$ with user u. The recommendation model is trained by minimizing loss, comparing labels y with predicted score $\hat{y}$ for the final prediction.

Symbol	Definition	Description
$\mathcal{U},\ \mathcal{I},\ \mathcal{R}$	$\mathcal{U} = \{u_1, u_2, ..., u_j\}$, $\mathcal{I} = \{i_1, i_2, ..., i_k\}$, $\mathcal{R} = \{(u, i, y) \mid u \in \mathcal{U}, i \in \mathcal{I}\}$	The set of users, items(videos), and user-item interactions
C_i	$C_i = \{c_1, c_2, \ \dots, c_l\}$	The clip set of videos
$C_u^{\mathbf{intrct.}}$	$C_u^{\mathbf{intrct.}} = \cup_{i=1}^{k} \{c_{i,1}^{\mathbf{intrct.}}, c_{i,2}^{\mathbf{intrct.}}, \ \dots, c_{i,l}^{\mathbf{intrct.}}\}$	The interacted clip set of users
$\mathbf{e},\ \mathbf{f}$	$\mathbf{e}_u, \mathbf{e}_i, \mathbf{f}_u^m, \mathbf{f}_i^m \in \mathbb{R}^d$, $m \in \mathcal{M} = \{v, a, t\}$	Embedding and Feature representation
$\hat{x}$	$\hat{x} = max_{\ c^{\mathbf{pos}},\, c^{\mathbf{nov}} \in C,\, c^{\mathbf{pos}} \neq c^{\mathbf{nov}}} Sim(c^{\mathbf{pos}},\ c^{\mathbf{nov}})$	Novelty score based on similarity
$\hat{\mathrm{y}}$	$\hat{\mathrm{y}} = \lambda \cdot score\_(u,i)^{ns} + (1\text{-}\lambda) \cdot score\_(u,i)^{ds}$	Preference label and predicted score

Proposed Method

Graphical Abstract

- The video features are processed through distinct paths corresponding to the adjacency matrices from the non-skip and delayed skip graphs, with each path employing a 2-layer GCN architecture, 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.

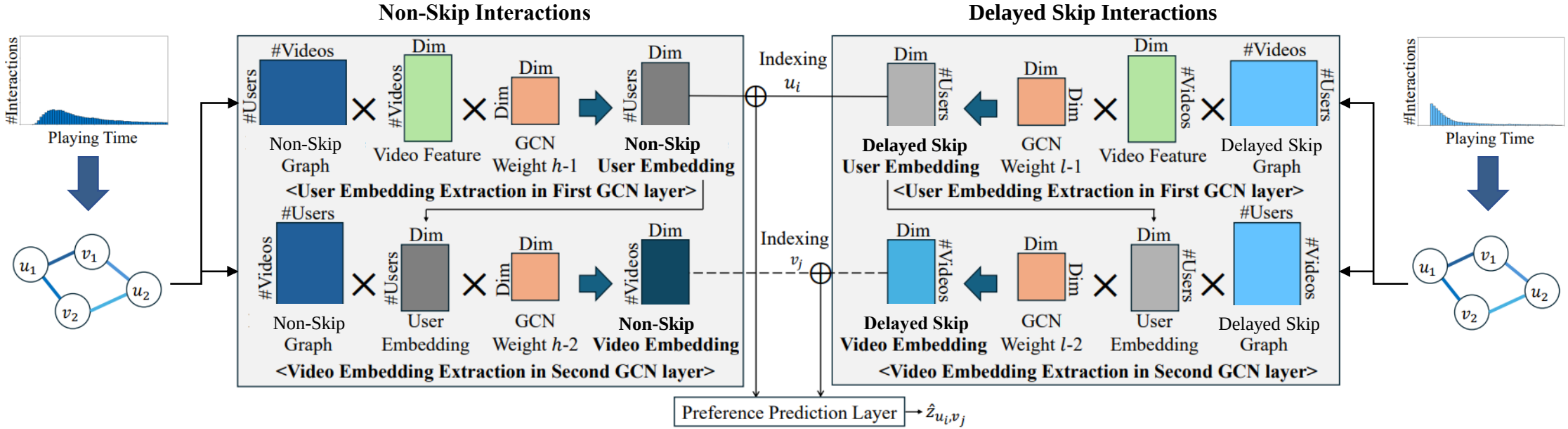


Figure 7. Proposed Method Architecture

Proposed Method

Details

- The total interactions are initially divided into non-skip interactions and skip behavior interactions based on duration/playing time, after which the skip behavior interactions are further divided into delayed skip interactions and immediate skip interactions.
- Non-skip interactions and delayed skip interactions each form individual adjacency graphs that are utilized in dual-path positive graph learning, while immediate skip interactions help the model learn preference differences between interactions through a ranking loss.

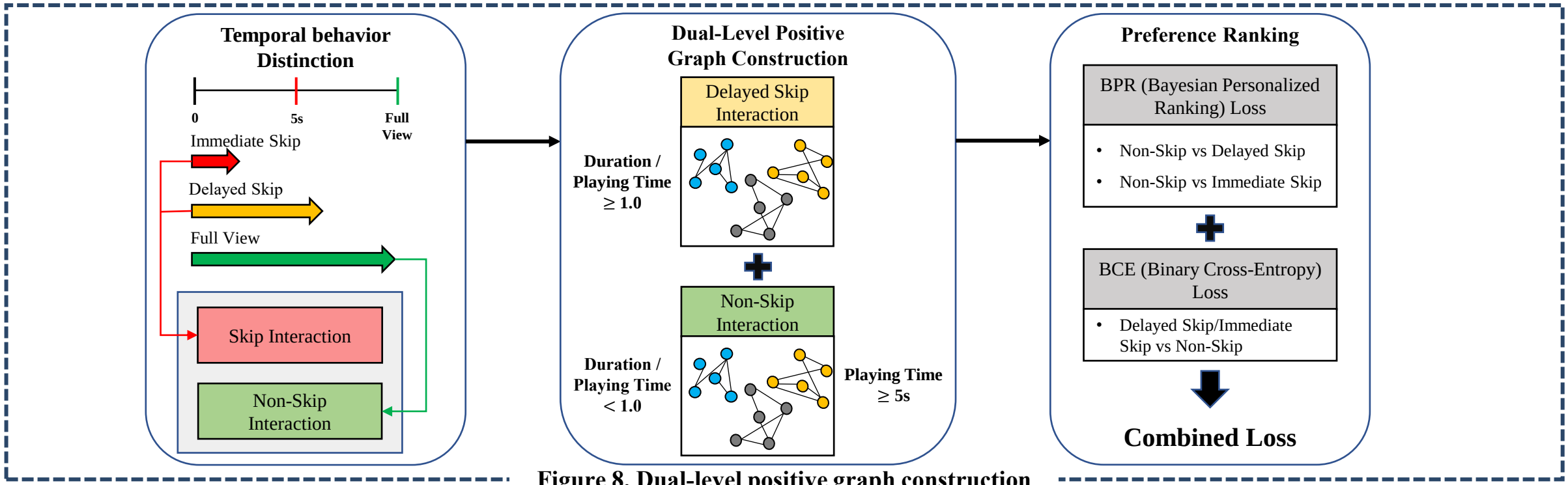


Figure 8. Dual-level positive graph construction

Experimental Settings

Dataset

Dataset	#Interactions	#Users	#Videos	V	A	T
KuaiRand-Pure (Random policy)	1,186,059	27,285	7,583	-	-	-
MicroVideo-A	342,694	12,739	58,291	128	-	-

Dataset Description

- The experiments were conducted with three real-world datasets collected from actual platforms to compare with previous studies.
- The datasets comprise interaction records and video characteristics, including user actions such as user ID, video ID, playback time, video duration, timestamp, and follow.

Evaluation Measures

- Precision@k: The ratio of relevant items among the top k items recommended by the model.
- Recall@k: The ratio of actual user-preferred items included in the top k items recommended by the model.
- MAP@k (Mean Average Precision): The ratio that considers both relevance and ranking position of recommended items up to position k.
- NDCG@k: The ratio of the top k recommended items that the model has ranked in the same order as the user actual preferences.

Experimental Settings

Cross Validation

- Number of Iterations: 10
- Data Split Ratio (Train: Validation: Test) = 0.6 : 0.2: 0.2

Hyperparameter Settings

Method	Batch Size	Max Epochs	Loss Function	Optimizer	Learning Rate
Proposed FRAME(2023) LightGT(2023) BM3(2023) DGRec(2023)	512	20	BPR Loss + BCE Loss	AdamW	1e-3

Experimental Results

Performance Comparison

- The proposed dual-graph-based micro-video recommender, designed to aggregate user preferences effectively, achieves higher evaluation metric scores than the comparison methods.

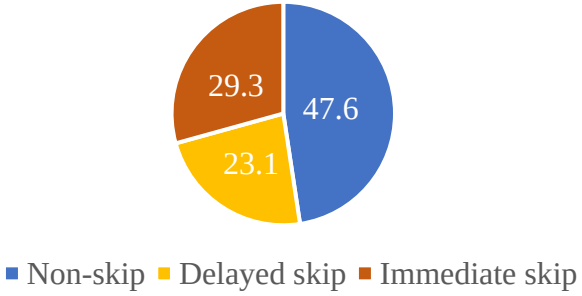
Dataset	**KuaiRand-Pure**								**MicroVideo-A**							
Metric	Precision@5	Recall@5	MAP@5	NDCG@5	Precision@10	Recall@10	MAP@10	NDCG@10	Precision@5	Recall@5	MAP@5	NDCG@5	Precision@10	Recall@10	MAP@10	NDCG@10
Proposed	**0.213 ± 0.002**	**0.712 ± 0.005**	**0.499 ± 0.005**	**0.577 ± 0.005**	**0.184 ± 0.000**	**0.857 ± 0.000**	**0.505 ± 0.003**	**0.613 ± 0.003**	**0.527 ± 0.004**	**0.838 ± 0.002**	**0.685 ± 0.003**	**0.777 ± 0.002**	**0.519 ± 0.004**	**0.995 ± 0.000**	**0.675 ± 0.003**	**0.783 ± 0.002**
FRAME	0.206 ± 0.002	0.677 ± 0.006	0.465 ± 0.004	0.543 ± 0.005	0.181 ± 0.001	0.833 ± 0.003	0.472 ± 0.003	0.582 ± 0.003	0.521 ± 0.006	0.829 ± 0.003	0.661 ± 0.006	0.757 ± 0.005	**0.519 ± 0.004**	**0.995 ± 0.000**	0.654 ± 0.005	0.766 ± 0.004
LightGT	0.164 ± 0.001	0.553 ± 0.000	0.318 ± 0.003	0.395 ± 0.002	0.164 ± 0.001	0.748 ± 0.001	0.338 ± 0.003	0.457 ± 0.003	0.521 ± 0.003	0.828 ± 0.004	0.681 ± 0.006	0.773 ± 0.005	**0.519 ± 0.004**	0.994 ± 0.000	0.670 ± 0.005	0.780 ± 0.004
BM3	0.165 ± 0.003	0.557 ± 0.006	0.327 ± 0.007	0.403 ± 0.007	0.165 ± 0.002	0.750 ± 0.006	0.347 ± 0.007	0.464 ± 0.006	0.522 ± 0.004	0.831 ± 0.001	0.676 ± 0.008	0.769 ± 0.006	**0.519 ± 0.004**	0.994 ± 0.000	0.666 ± 0.007	0.776 ± 0.006
DGRec	0.164 ± 0.001	0.553 ± 0.000	0.324 ± 0.003	0.400 ± 0.003	0.165 ± 0.001	0.750 ± 0.002	0.344 ± 0.003	0.462 ± 0.003	0.522 ± 0.003	0.830 ± 0.002	0.682 ± 0.004	0.774 ± 0.003	0.518 ± 0.004	0.994 ± 0.000	0.671 ± 0.004	0.781 ± 0.003

Experimental Results

In-depth Analysis

- The experiments conducted using the MicroVideo-A dataset for the proposed model aimed to investigate the impact of separate graph modeling for delayed skips on performance.
- The fine-grained skip behavior modeling achieved performance improvements across all metrics compared to the existing positive/negative classification method, demonstrating that separate modeling effectively contributes to enhanced accuracy.

Skip Behavior Pattern Analysis



- **Non-skip:**
  Sample count: 37,654 / Avg Watch time: 28.4 seconds
- **Delayed skip:**
  Sample count: 18,232 / Avg Watch time: 15.1 seconds
- **Immediate skip:**
  Sample count: 23,067 / Avg Watch time: 2.2 seconds

MicroVideo-A	Positive/Negative Distinction	Non-skip/Delayed skip/ Immediate skip distinction	Improvement
Precision@5	0.508	**0.570**	+ 0.062
Recall@5	0.188	**0.191**	+ 0.003
NDCG@5	0.811	**0.868**	+ 0.057
Precision@10	0.413	**0.485**	+ 0.072
Recall@10	0.312	**0.316**	+ 0.004
NDCG@10	0.799	**0.847**	+ 0.048

Overall Performance Comparison

Experimental Results

In-depth Analysis

- Conducting recommendations using viewing histories of users with high delayed skip ratios in a specific iteration as input, examined the outputs of both the positive/negative Distinction and the non-skip/delayed skip/immediate skip Distinction method.
- Evaluation results based on viewing records excluded from training revealed items that failed in prediction when considered negative in the positive/negative classification approach, but succeeded when classified as delayed skip and modeled separately.

Delayed skip user cluster

- **Viewing Pattern**:
  Non-skip:19% / Delayed skip: 63% / Immediate skip: 28%
- **Average skip delayed time**: 52.3 seconds

Video	Delayed skip Interaction/Total Interaction(%)
video_1205	26/35 (75.6%)
video_1456	14/20 (73.7%)
video_1789	29/43 (67.5%)
video_2001	22/34 (65.3%)
video_2156	28/44 (63.7%)
⋮	

Viewing history

Rank	Recommend Item	Confusion matrix	Distinction Type
1	video_1401	False Positive	Negative
2	video_2704	False Positive	Negative
3	video_3805	True Positive	Positive
4	video_4403	False Positive	Negative
5	video_4504	False Positive	Negative

Positive/Negative Distinction Recommend item List

Rank	Recommend Item	Confusion matrix	Distinction Type
1	video_1401	True Positive	Delayed skip
2	video_2704	False Positive	Delayed skip
3	video_3805	True Positive	Non-skip
4	video_4403	False Positive	Immediate skip
5	video_4504	True Positive	Delayed skip

Non-skip/Delayed skip/Immediate skip Distinction Recommend item List

Conclusion and Contribution

- This study proposes a dual-graph-based micro-video recommender system that effectively leverages fine-grained skip behaviors to improve recommendation accuracy.
- While conventional approaches oversimplify skip behavior modeling by categorizing interactions solely as positive or negative based on whether skipping occurs, our method addresses this limitation by refining interaction signals into three distinct categories.
- The experimental results demonstrate that our proposed approach significantly outperforms three state-of-the-art baseline methods across eight evaluation measures on two public micro-video datasets

Contribution

- Journal
  - Survey on Replay-based Continual Learning and Empirical Validation on Feasibility in Diverse Edge Devices using Representative Method, Mathematics, 2025
- In International Conferences
  - Automatic Playlist Generation: A Comprehensive Crux of a Methodology for Technology Trend, ICCE, 2025
  - (Submitted) TMV-GCN: Temporal Multi-View Graph Convolution Networks for Micro-Video Recommendation
- In Domestic Conferences
  - Enhancing Representation Learning through Transformer with Positional Embedding for Automatic Playlist Continuation, IEIE, 2025

Thank you

Reference

[1] S. Liu et al. "User Conditional Hashtag Recommendation for Micro-Videos." *2020 IEEE Int. Conf. Multimed*. Expo, London, UK, Jul. 2020, pp. 1-6.

[2] A. Chaudhary et al. "“Are you still watching?”: exploring unintended user behaviors and dark patterns on video streaming platforms." in *Proc. ACM Design. Interac. Sys. Conf.*, Australia, 2022, pp. 776-791.

[3] Z. Fu et al. "Deep Learning Models for Serendipity Recommendations: A Survey and New Perspectives." *ACM Comput. Surv.*, vol. 56, no. 19, Aug. 2023, pp. 1-26.

[4] M. Wilhelm et al. "Practical Diversified Recommendations on YouTube with Determinantal Point Processes." in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manag.*, Torino, Italy, 2018, pp. 2165-2173.

[5] D. Klug et al. "Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm." in *Proc. 13th ACM Web Sci. Conf.*, United Kingdom, 2021, pp. 84-92.

[6] M. Boeker et al. " An Empirical Investigation of Personalization Factors on TikTok." in *Proc. ACM Web Conf.*, Lyon, France, 2022, pp. 2298-2309.

[7] S. Yu et al. "Learning Fine-grained User Interests for Micro-video Recommendation." in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Taipei, Taiwan, 2023, pp. 433-442.

[8] Y. Wei et al. "LightGT: A Light Graph Transformer for Multimedia Recommendation." in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr*., Taipei, Taiwan, 2023, pp. 1508-1517.

[9] Zhou. Xin et al. "Bootstrap latent representations for multi-modal recommendation." In *Proceedings of the ACM web conference 2023*, pp. 845-854.

[10] Yang. L et al. "Dgrec: Graph neural network for recommendation with diversified embedding generation." In Proceedings of the sixteenth ACM international conference on web search and data mining, 2023, pp. 661-669.