A Multi-modal Modeling Framework for Cold-start Short-video Recommendation

Gaode Chen
Kuaishou Technology
Beijing, China
chengaode19@gmail.com

Ruina Sun
Kuaishou Technology
Beijing, China
sunruina@kuaishou.com

Yuezihan Jiang*
Kuaishou Technology
Beijing, China
jiangyuezihan@kuaishou.com

Jiangxia Cao
Kuaishou Technology
Beijing, China
jiangxiacao@gmail.com

Qi Zhang*
Kuaishou Technology
Beijing, China
zhangqi38@kuaishou.com

Jingjian Lin
Kuaishou Technology
Beijing, China
linjingjian@kuaishou.com

Han Li
Kuaishou Technology
Beijing, China
lihan08@kuaishou.com

Kun Gai
Kuaishou Technology
Beijing, China
yuyue06@kuaishou.com

Xinghua Zhang
Institute of Information Engineering,
Chinese Academy of Sciences
Beijing, China
zhangxinghua@iie.ac.cn

ABSTRACT

Short video has witnessed rapid growth in the past few years in multimedia platforms. To ensure the freshness of the videos, platforms receive a large number of user-uploaded videos every day, making collaborative filtering-based recommender methods suffer from the item cold-start problem (e.g., the new-coming videos are difficult to compete with existing videos). Consequently, increasing efforts tackle the cold-start issue from the content perspective, focusing on modeling the multi-modal preferences of users, a fair way to compete with new-coming and existing videos. However, recent studies ignore the existing gap between multi-modal embedding extraction and user interest modeling as well as the discrepant intensities of user preferences for different modalities. In this paper, we propose M3CSR, a multi-modal modeling framework for cold-start short video recommendation. Specifically, we preprocess content-oriented multi-modal features for items and obtain trainable category IDs by performing clustering. In each modality, we combine modality-specific cluster ID embedding and the mapped original modality feature as modality-specific representation of the item to address the gap. Meanwhile, M3CSR measures the user modality-specific intensity based on the correlation between modality-specific interest and behavioral interest and employs pairwise loss to further decouple user multi-modal interests. Extensive experiments on four real-world datasets demonstrate the superiority of our proposed model. The framework has been deployed on a billion-user scale short video application and has shown improvements in various commercial metrics within cold-start scenarios.

*Corresponding author.

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688098

CCS CONCEPTS

• **Information systems** → Recommender systems.

KEYWORDS

Recommender System, Cold-start, Multi-modal Modeling

ACM Reference Format:
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 *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3640457.3688098

1 INTRODUCTION

Short video applications like TikTok and Kwai have grown rapidly in recent years. Tens of millions of short videos are being generated by users every day, which has greatly improved the richness and freshness of content ecology. Meanwhile, the personalized recommender systems [11] play a vital role in accurately recommending appropriate videos for users to alleviate information overload.

Unfortunately, the item cold-start problem [1, 39] may occur when the recommendation model faces a large volume of new videos released every day. On the one hand, mainstream methods such as collaborative filtering (CF) algorithms [38] require historical user-item interactions to learn a meaningful ID embedding for each entity. Concretely, for a large amount of new emerging videos with limited interactions, their embeddings are insufficiently trained, resulting in new videos that may miss the opportunity to be recommended or be recommended to inappropriate users. On the other hand, an unbalanced percentage of cold-start videos in the total can cause the model to overweight the majority of well-trained videos, which is not conducive to the sustainable development of the platform content. Thus, the cold-start problem has become a crucial obstacle for online recommendation.

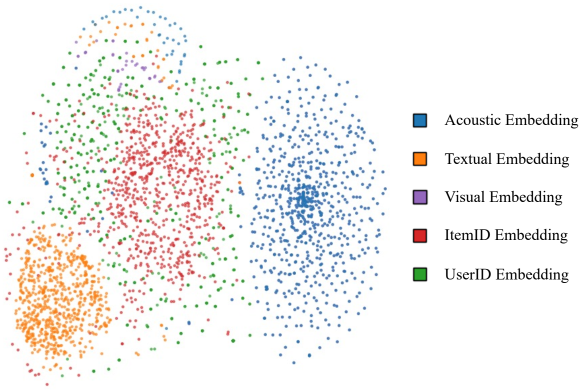


Figure 1: t-SNE visualization of content and ID embeddings.

Based on the rich multi-modal content of short videos (e.g., visual, textual, and acoustic), considerable efforts have been made to solve the cold-start problem. For example, VBPR [10], DeepStyle [21], and ACF [4] extend the vanilla CF framework by incorporating multi-modal contents as side information in addition to ID embedding of items. The performance lift comes from incorporating the content information like text caption, image cover, and soundtrack representations. However, in real-world industrial applications, the information mentioned above has long been incorporated to accurately portray an item, and the **online models still mainly rely on high-quality ID embedding**, which confines the expressiveness of content features and lacks generalization. Alternatively, the approach we seek is expected to recommend cold start videos to appropriate users from a content perspective by modeling the multi-modal interests of users and based on the multi-modal features of cold start videos. Such modeling is more general and can improve the distribution efficiency of cold-start videos. We argue that two challenges remain, which harm the multi-modal modeling process and cause sub-optimal performance.

A gap exists between the multi-modal embedding extraction and user interest modeling. In practical industrial settings, the multi-modal embeddings of videos are pre-trained with mature encoders such as BERT [16] for textual. Due to the high online real-time requirements, these multi-modal embeddings are usually treated as *non-trainable* embeddings to model user interests, rather than optimizing the content encoders end-to-end. Furthermore, multi-modal embedding extraction is a module tailored for upstream tasks such as video classification. Thus, it is a content-oriented and non-personalized process but not to the user preferences. However, user interest modeling is focused on understanding interactions and is essentially a behavior-oriented and personalized process. In Figure 1, we visualize the distribution of the three modalities and personalized ID embeddings from a real-world short-video platform by reducing their dimension to two with t-SNE [31]. The significant *distribution difference* between content and ID embeddings poses a challenge in modeling user interests. In general, these gaps restrict the positive impact of multi-modal information.

The discrepant intensity of user preferences across different modalities. Recently, a few attempts [20, 29, 35, 37] have been made to model user preferences at fine-grained modality levels. For instance, MMGCN [35] constructs modality-specific user-item interaction graphs to model user preference for each modality. Despite their effectiveness, previous attempts fail to explicitly model that users tend to place different emphasis on different modalities. In our intuition, some users are easily attracted by the cover image of the short videos, but may also be turned away by its poor soundtrack. There may be some users are more focus on the theme of the video, i.e., the text caption, regardless of the frame or soundtrack of the video. In a nutshell, user preference for short videos depends not only on the match between the video content and user multi-modal interests but also on their modality-specific interest intensity.

To address the above issues, we propose a novel and practical **M**ulti-**m**odal **M**odeling framework for **C**old-start **S**hort-video **R**ecommendation, namely M3CSR. Our framework is designed as a dual-tower architecture [14], which is the mainstream structure used by online recommender systems in the retrieval stage. Before training the model, we conduct a pre-processing showcasing the generation process of multi-modal embeddings. Additionally, we utilize the K-means algorithm [18] to obtain stable cluster centers from the overall multi-modal embeddings of videos and assign a cluster center ID to each video. In M3CSR, we establish a separate trainable embedding table for the multi-modal cluster ID in each modality. To address the gap caused by the *non-trainable* characteristic and *distribution difference* of content embeddings, we design a **Modal Encoder** to obtain modality-specific representation for each item, which combines trainable modality-specific cluster ID embedding and a mapped representation of the original modality features. Our encoder elegantly evolves from non-trainable content embeddings to trainable cluster ID embeddings, expanding the modeling space for co-occurrence relation learning of multiple modalities and narrowing the gap between multi-modal embedding extraction and user interest modeling.

Meanwhile, for items in the user historical sequence, we can obtain modality-specific sequential representations for the user based on the Modal Encoder in each modality. Furthermore, for modality-specific sequential representations, we utilize user ID embedding as a query and employ a multi-head attention mechanism [32] to model user modality-specific interest. More importantly, we develop a **Multi-modal Interest Intensity Learning** network to measure the genuine tastes of users across different modalities. Concerns about over-optimization, we also leverage pairwise loss to learn more decoupled multi-modal interest representations.

Moreover, we argue that ID embedding is inaccurate when the target video is in a cold-start state and content embeddings should exert a more significant influence. Thus, M3CSR learns a gating network based on the popularity features of the video, such as view counts, to control the weights of video-side ID embedding and content embeddings. At the top of the user- and item- tower, we concatenate the behavioral and content representations respectively as the final output. This ensures that whether using the inner product during training or Approximate Nearest Neighbor (ANN) during serving, the bilateral matching effects between different components are cumulative rather than interfering with each other, minimizing the impact of distribution differences.

To summarize, the key contributions are as follows:

- We highlight the gap between multi-modal embedding extraction and user interest modeling in existing methods. M3CSR has overcome the dilemma by introducing impressively optimizable multi-modal cluster IDs and a well-designed Modal Encoder.
- We propose that users tend to demonstrate various intensities across different modalities, and we utilize pairwise loss to maintain the distinction of user preferences among these modalities. Subsequent experimental results validate the necessity of addressing this problem.
- Extensive offline experiments on four real-world datasets validate the effectiveness of the proposed method. Currently, M3CSR has been deployed on a billion-user scale short video application, yielding significant improvement on a series of commercial metrics in cold-start scenarios.

2 RELATED WORKS

Cold-start Recommendation. Cold-start problem is one of the main challenges in recommender systems. The common solution to this issue can be categorized into two types, namely content-based and transfer learning based methods. The first type of methods [23, 30, 33] aims to exploit content information, such as item attributes, to enhance the recommendation performance. These methods are proposed with the assumption that if a user likes a item, it is very likely that she/he will prefer other content-similar items. By building this user content preference, the content-based filtering is able to make cold-start item recommendation without requiring any behavior logs for new items. For example, LCE [26] exploits items' properties and past user preferences by a local collective embedding learning method. Another way to alleviate the cold-start problem is to transfer knowledge from other domains, such as cross-domain recommendation [36], transfer learning methods [27], and meta-learning methods [6]. In this work, we propose a new short-video recommender system that addresses the challenge faced by existing methods in dealing with the massive newly released cold-start short videos daily. Our method is content-based and models the genuine distribution of user multi-modal interests from a content perspective, enabling effective generalization.

Multi-modal Recommendation. Many efforts [19, 28] have been devoted to enhancing recommender systems by incorporating multi-modal content. One representative early study VBPR [10] extends matrix factorization to integrate both ID embeddings and visual features of items. To improve the user-item relation modeling with multi-modal content, attention mechanisms are used in ACF [4] and VECF [5] to capture complex user preference. In recent years, a few works have explored capturing user fine-grained preferences on different modalities. For example, MMGCN [35] tries to model the user preferences on the modal-specific user-item bipartite graph. However, these methods directly utilize multi-modal features as side information or model user preferences uniformly across modalities. Our model further optimizes the modeling of multi-modal features and learns the interest intensity of users to different modal content.

3 PRELIMINARY

Definitions of notations. Let $\mathcal{U}$, $\mathcal{I}$ denote the set of users and items (short videos), respectively. Each user $u \in \mathcal{U}$ is associated with a set of items $\mathcal{I}^u$ with positive feedbacks which indicate the preference score $y_{ui} = 1$ for $i \in \mathcal{I}^u$. $\mathbf{e}_u^{id}$, $\mathbf{e}_i^{id} \in \mathbb{R}^d$ is the input ID embedding of u and i, respectively, where d is the embedding dimension. Besides user-item interactions, multi-modal features are offered as content information of items. We denote the modality features of item i as $\mathbf{e}_i^m \in \mathbb{R}^{d_m}$, where d_m denotes the dimension of the m-modal feature, $m \in \mathcal{M}$ is the modality, and $\mathcal{M}$ is the set of modalities. In this paper, we consider visual, textual, and acoustic modalities denoted by $\mathcal{M} = \{v, t, a\}$. Please kindly note that our method is not fixed to the three modalities.

Task Formulation. We focus on addressing the item (short-videos) cold-start problem, which is the problem that new items have no or rare prior events. We formulate our multi-modal recommender system that captures user-item relations with modality-aware user preference learning. In particular, given the pair of (u, i), $u \in \mathcal{U}$ and $i \in \mathcal{I}$, our task is to learn a function that forecasts how likely the item will be adopted by the user, i.e. $\hat{y}_{ui}$.

4 METHODOLOGY

4.1 Overview

Figure 2 shows the architecture of M3CSR, which is a dual-tower architecture, that is, there is no feature cross or structure cross between user- and item-side modeling. This is also the mainstream structure used by the industrial recommender system in the retrieval stages. Additionally, if the recommendation model is to be deployed online, user and item embeddings can be pre-computed and indexed using an ANN search system, such as FAISS [15], so that we can retrieve top-N relevant items efficiently within the high real-time demands at serving.

In terms of the gap between multi-modal embedding extraction and user interest modeling, we design a **Modal Encoder** (Figure 3) to obtain the modality-specific representation of the item based on its modality-specific original embedding and modality-specific cluster ID embedding. The above multi-modal features are obtained in the preprocessing procedure. In the user-side tower, on the one hand, we can obtain the user behavioral interest by modeling the user behavior sequence. On the other hand, we utilize the Modal Encoder to convert the user behavior sequence to obtain sequence representations in each modality. Furthermore, we model the high-order connectivity between users and short-videos in the above sequences for each modality to capture user preference on modality-specific content. We explicitly scale user tastes for multi-modal contents in the **Multi-modal Interest Intensity Learning** network, and utilize pairwise loss to further disentangle user interests at the granularity of modality. In the item-side tower, in addition to the item ID embedding, we also use the Modal Encoder to obtain the content embeddings of the item in each modality. It is noteworthy that the popularity features of the item is applied to control the effect of the content embeddings when the ID embedding is not learned accurately enough in its cold start state.

4.2 Preprocessing

Before recommender model training, we have to briefly describe the techniques utilized for extracting embedding vectors from different modalities like visual, textual, and acoustic. Besides, we also preprocess multi-modal clustering centers for further learning of modality-specific co-occurrence relationships.

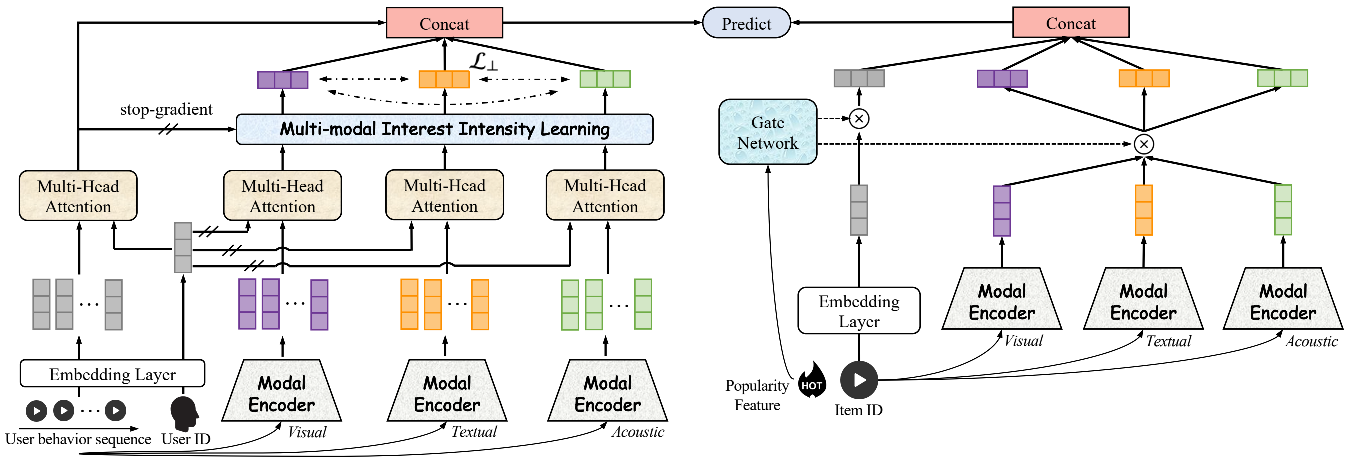


Figure 2: The overview of our proposed M3CSR: Left is the user-side tower and right is the item-side tower.

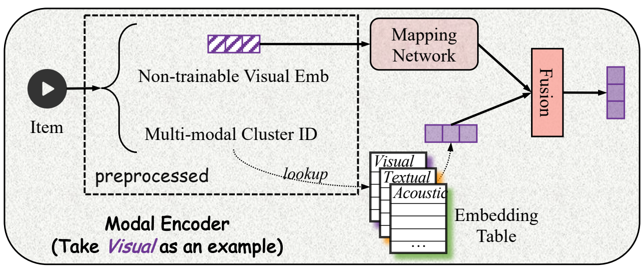


Figure 3: The structure of Modal Encoder.

Multi-modal Embedding. For the processing of offline datasets, we directly follow the results of previous work[23, 33]. For example, they employed ResNet[9] for visual modality to process each frame of the video and then perform average aggregation, employed Sentence-BERT[24] for textual modality to process the summary sentences of the video, and employed VGGish[12] for the acoustic modality to process the soundtrack of the video. In real-world application scenarios, researchers also utilize these pre-trained encoders. However, this part is usually completed by the Middle Platform which stores the results of multi-modal embeddings for online recommender models. This will cause models to only obtain the original multi-modal embeddings of videos in a non-trainable way, and we cannot optimize the modality encoder end-to-end.

Multi-modal Cluster ID. We create trainable IDs for the non-trainable multi-modal embeddings, i.e. multi-modal cluster IDs. We consider that cluster ID has the advantage of being more uniform and fine-grained than categories or tags. It can help us capture collaborative signals in different modalities to alleviate the gap. For video $i \in \mathcal{I}$, we concatenate its modal embeddings as its overall content representation, i.e., $\mathbf{e}_i^c = concat(\mathbf{e}_i^v, \mathbf{e}_i^t, \mathbf{e}_i^a)$. Then, we directly use the K-means algorithm [18] to cluster the content representations of all videos, obtain several cluster centers, and assign the nearest cluster center to each video i. We construct an embedding table of cluster IDs for each modality and use the look-up operation to convert the cluster ID of item i into the low-dimensional latent space of different modalities, i.e. $\mathbf{c}_i^m \in \mathbb{R}^d$, $m \in \mathcal{M}$. In contrast, in our application scenario, we performed K-means clustering based on at least 10 million videos and obtained 1,000 cluster centers. We can consider the cluster centers to be stable because the content-oriented embedding algorithms are non-personalized. Therefore, new videos released every day will be assigned to one of these 1,000 centers based on the nearest principle.

4.3 Modal Encoder

After non-personalized feature extraction, the multi-modal embedding of the video is denoted as $\mathbf{e}_i^m \in \mathbb{R}^{d_m}$, $m \in \mathcal{M}$. However, there is a large distribution difference between these multi-modal embeddings and other ID embeddings. Another point is that in real-world industrial environment, these multi-modal embeddings are usually used as external embedding that cannot be optimized.

Thus, we attempt to solve the above issues from two aspects. First, we apply a dense network to learn the mapping relationship from content space to behavior space to reduce the impact of feature distribution differences. Secondly, we perform clustering preprocessing based on multi-modal embedding, assign a trainable cluster ID embedding $\mathbf{c}_i^m \in \mathbb{R}^d$, $m \in \mathcal{M}$ to the item $i \in \mathcal{I}$. We map the modality-specific content embedding $\mathbf{e}_i^m$ and concatenate it with the modality-specific cluster ID embedding $\mathbf{c}_i^m$ to reverse the untrainable dilemma, as the modality-specific representation $\mathbf{h}_i^m \in \mathbb{R}^{2d}$ of item i.

$$\mathbf{h}_i^m = \text{Encoder}\,(i)\,, m \in \mathcal{M} \tag{1}$$

$$\text{Encoder}\,(i) = concat\left(\mathbf{w}_m^\top \mathbf{e}_i^m + b_m, \mathbf{c}_i^m\right), m \in \mathcal{M} \tag{2}$$

where $\left[\mathbf{w}_m^\top; b_m\right]$ are parameters of mapping network specific to the modality m. It is worth noting that different modalities have different mapping networks with separate parameters and the trainable embedding space for modality-specific cluster IDs is shared between the user- and the item-side.

4.4 User-side Tower

4.4.1 Sequence Modeling. The modeling of collaborative relationships is crucial in recommendation systems, and ID embedding still dominates personalized recommendations. Despite explicitly modeling user preferences for video content, we are still of the opinion that content embedding cannot solve the data sparsity and cold-start problems alone. We take user $u \in \mathcal{U}$ as an example to demonstrate the process of sequence modeling. Given a fixed-length user behavior sequence $s_u = <i_1^u, i_2^u, ..., i_n^u>$, where n represents the maximum sequence length we consider. The embedding look-up operation converts the IDs of items in the sequence into a unified low-dimension latent space, we can obtain its embedding $\mathbf{H}_u = <\mathbf{e}_{i_1^u}^{id}, \mathbf{e}_{i_2^u}^{id}, ..., \mathbf{e}_{i_n^u}^{id}>$. After the embedding layer, we adopt the multi-head attention [32] to capture the user behavioral interest $\mathbf{h}_u \in \mathbb{R}^d$ in the sequence. The computation is as follows:

$$\mathbf{h}_u = \text{MH}\left(Q = \mathbf{e}_u^{id}, K = \mathbf{H}_u, V = \mathbf{H}_u\right) \tag{3}$$

4.4.2 User Multi-modal Interest Learning. Multi-modal information is the dominant presentation of the item and it directly engages with users. Therefore, it contains abundant user preference-related clues that differ from the collaborative relationships in the interactions. In the flourishing Internet age, there are thousands of new short videos being published every second. These new short videos suffer from cold-start problems due to their sparse interaction data in the early period, which means that their ID embeddings are not learned accurately enough, making it difficult for the recommendation system to distribute them to the appropriate users. Therefore, we need to consider enhancing user representations and further capturing user fine-grained preferences on different modalities. In this way, we can make recommendations in a more generalized way based on the content of new short videos and user content preferences, accumulate valuable interactive information for new videos, and help solve the cold start problem.

Based on the user sequence s_u, we can utilize the Modal Encoder to convert the item IDs into content embeddings $\mathbf{H}_u^m$ in different modalities, i.e., $\mathbf{H}_u^m = \text{Encoder}(s_u)$, $m \in \mathcal{M}$. We still utilize user ID embedding as a query to model multi-modal sequence through the multi-head attention mechanism and obtain fine-grained modality-specific interest $\mathbf{h}_u^m \in \mathbb{R}^d$ of user u.

$$\mathbf{h}_u^m = \text{MH}\left(Q = \text{sg}\left(\mathbf{e}_u^{id}\right), K = \mathbf{H}_u^m, V = \mathbf{H}_u^m\right), m \in \mathcal{M} \tag{4}$$

where sg denotes the stop-gradient operator.

In addition, users usually have discrepant preference intensities for different modalities. When we recommend new videos that users may be interested in, except considering the degree of matching between the user content preference and the content features of the new video, we should also take into account the user's own sensitivity to different modalities. Homogenizing or unifying multi-modal channels is insufficient to recognize the different importance of modalities, hampering information propagation and leading to suboptimal representations. In this paper, we argue that a user modality-specific intensity is jointly determined by his/her behavioral preference and modality-specific preference. Modality-specific preference is approximately close to the user behavioral preference, and the user is more sensitive to the content of this modality. Specifically, we concatenate the user behavioral interest $\mathbf{h}_u$ with each user modality-specific interest $\mathbf{h}_u^m$ to learn the intensity factor λ_m of each modality through the Multi-modal Interest Intensity network. Finally, we perform softmax [13] on the intensity factors of all modalities.

$$\lambda_m = \mathbf{w}_{int}^{\top} concat\left(\text{sg}\left(\mathbf{h}_u\right), \mathbf{h}_u^m\right), m \in \mathcal{M} \tag{5}$$

$$\overline{\lambda}_m = \frac{\exp\left(\lambda_m/\tau\right)}{\sum_{j \in \mathcal{M}} \exp\left(\lambda_j/\tau\right)}, m \in \mathcal{M} \tag{6}$$

where $\mathbf{w}_{int}^{\top}$ are trainable weights of the Multi-modal Interest Intensity Learning network, exp is the exponential operation with the base e, and τ is the temperature coefficient.

After that, we use the modality-specific normalized intensity factors $\overline{\lambda}_m$ to get the weighted modality-specific interest $\widetilde{\mathbf{h}}_u^m$, reflecting the scaling of our model to the user multi-modal tastes.

$$\widetilde{\mathbf{h}}_u^m = \overline{\lambda}_m \mathbf{h}_u^m, m \in \mathcal{M} \tag{7}$$

We are concerned that the user multi-modal preferences will be optimized toward his/her behavioral preference, resulting in no differentiation. Thus, we minimize the square of dot product of pairwise vectors among the two modality-specific interests.

$$\mathcal{L}_{\perp} = \left\|\widetilde{\mathbf{h}}_u^v \cdot \widetilde{\mathbf{h}}_u^a\right\|^2 + \left\|\widetilde{\mathbf{h}}_u^v \cdot \widetilde{\mathbf{h}}_u^t\right\|^2 + \left\|\widetilde{\mathbf{h}}_u^a \cdot \widetilde{\mathbf{h}}_u^t\right\|^2 \tag{8}$$

In the end, we concatenate the user behavioral interest with the weighted multi-modal interests as the user final representation.

$$\mathbf{h}_u^{final} = concat\left(\mathbf{h}_u, \widetilde{\mathbf{h}}_u^v, \widetilde{\mathbf{h}}_u^t, \widetilde{\mathbf{h}}_u^a\right) \tag{9}$$

4.5 Item-side Tower

After the release of a new video, it may go through multiple phases. In the early period, due to sparse interaction behavior, its ID embedding might not be sufficiently accurate. With the increasing number of views on new videos, their ID embeddings will be gradually optimized by the model, allowing the recommendation system to identify the user groups interested in them. We aim to design a gating mechanism that adjusts the weights of ID embedding and content embeddings based on video popularity information, such as view counts. This mechanism seeks to amplify the influence of content embeddings during the cold-start period of a video, mitigating the impact of its inaccurate ID embedding. We represent $\mathbf{e}_i^{pop} \in \mathbb{R}^d$ as the embedding of the popularity information for item i. In offline datasets, we count the interactions for each item and bin them based on these counts. Each item is assigned to the corresponding bin, and the bin ID is converted into a trainable embedding representation $\mathbf{e}_i^{pop}$. In real-world industrial settings, the view count of a video is directly accessible as an attribute feature. We implement this gating mechanism using a dense network, utilizing a sigmoid activation function to obtain weights δ ranging from 0 to 1.

$$\delta = sigmoid\left(\mathbf{w}_{pop}^{\top} \mathbf{e}_i^{pop} + b_{pop}\right) \tag{10}$$

where $[\mathbf{w}_{pop}^{\top}; b_{pop}]$ are parameters of the gate network.

At last, we get item final representation as follow:

$$\mathbf{h}_i^{final} = concat\left(\left(1 - \delta\right) \mathbf{e}_i^{id}, \delta \mathbf{h}_i^v, \delta \mathbf{h}_i^t, \delta \mathbf{h}_i^a\right) \tag{11}$$

Table 1: Statistics of experimented datasets with multi-modal item Visual(V), Acoustic(A), Textual(T) contents.

Dataset	Amazon				Tiktok			Allrecipes	
	Sports		Baby						
Modality	V	T	V	T	V	A	T	V	T
Embed Dim	4096	1024	4096	1024	128	128	768	2048	20
User	35,598		19,445		9,319			19,805	
Item	18,357		7,050		6,710			10,067	
Interactions	256,308		139,110		59,541			58,922	
Sparsity	99.961%		99.899%		99.904%			99,970%	

4.6 Model Prediction & Optimization

With the final embeddings, our M3CSR model makes predictions on the unobserved interaction between user u and item i through $\hat{y}_{ui} = \left[\mathbf{h}_u^{final}\right]^\top \cdot \mathbf{h}_i^{final}$. To predict the interaction between the users and short-videos, we apply Bayesian Personalized Ranking (BPR) [25], which is a well-known pairwise ranking optimization framework, as the learning model. In particular, we model a triplet of one user and two short-videos, in which one of the short-videos is observed and the other one is not, formally as

$$\mathcal{L}_{BPR} = \sum_{y \in D_{train}} -\log\left(\text{Sigmoid}\left(\hat{y}_{u,i_p} - \hat{y}_{u,i_n}\right)\right) \tag{12}$$

i_p, i_n denotes the positive and negative samples for user u, D_{train} is the training set.

Formally, the jointly optimized objective is given (α, β are hyperparameters for loss term weighting):

$$\mathcal{L} = \mathcal{L}_{BPR} + \alpha \cdot \mathcal{L}_{\perp} + \beta \cdot \|\theta\|^2 \tag{13}$$

where the last term in $\mathcal{L}$ is the weight-decay regularization.

5 EXPERIMENTS

5.1 Experimental Settings

5.1.1 Dataset. Following [33], we conduct offline experiments on four real-world multi-modal recommendation datasets [1], i.e., Amazon-Sports, Amazon-Baby, Tiktok, and Allrecipes. Data statistics with multi-modal feature embedding dimensionality are reported in Table 1.

- **Amazon.** We adopt two benchmark datasets from Amazon with two item categories Amazon-Baby and Amazon-Sports. In those datasets, textual feature embeddings are generated via Sentence-Bert [24] based on the extracted text from the product title, description, brand, and categorical information. The product images are used to generate 4096-d visual feature embeddings of items.
- **TikTok.** This data is collected from TikTok platform to log the viewed short-videos of users. The multi-modal features are visual, acoustic, and title textual features of videos. The textual embeddings are also encoded with Sentence-Bert.
- **Allrecipes.** This dataset comes from one of the largest food-oriented social network platform by including 52,821 recipes in 27 different categories. For each recipe, its image and ingredients are considered as the visual and textual features. Following the setting in [7], 20 ingredients are sampled for each recipe.

[1] All datasets are publicly available at https://github.com/HKUDS/MMSSL/tree/main.

5.1.2 Evaluation Protocols. For each dataset, we used the ratio 8:1:1 to randomly split the historical interactions of each user and constituted the training set, validation set, and testing set. Moreover, following the widely-used evaluation metrics [2, 3, 34], we adopted Precision@N, Recall@N, and Normalized Discounted Cumulative Gain (NDCG@N) to evaluate the performance of methods. By default, we set N = 20 and reported the average values of the three metrics for all users in the testing set.

5.1.3 Baseline Methods. We compare M3CSR with the following state-of-the-art multi-modal recommender systems.

- **VBPR** [10]. Such model integrates the content features and ID embeddings of each item as its representation and uses the matrix factorization (MF) framework to reconstruct the historical interactions between users and items.
- **MMGCN** [35]. It applies three nonlinear GCNs to perform message passing on the user-item graphs that hold data of different modalities, respectively, so as to learn fine-grained modality-specific user preferences.
- **LATTICE** [37]. It proposes learning item-item structures for each modality and aggregating multiple modalities to obtain latent item graphs.
- **Siamese** [23]. This model concatenates visual, textual, acoustic, and meta-data embeddings to represent video, and introduces a Siamese-based network to predict similarities between user embedding and video embedding. It fully leverages multi-modal content embeddings without resorting to ID embeddings.
- **MMSSL** [33]. This is the most competitive method that derives self-supervision signals by effectively learning modality-aware user preferences and cross-modal dependencies to alleviate data sparsity issues. It introduces a modality-aware interactive structure learning paradigm, aiming to characterize the interdependence between the collaborative view and the multi-modal semantic view. Additionally, it introduces a cross-modal contrastive learning approach to jointly preserve the inter-modal semantic commonality and user preference diversity.

5.1.4 Implementation Details. We implement our method in PyTorch [22] and set the embedding dimension d fixed to 64 for all models. We optimize all models with the Adam [17] optimizer, where the batch size is fixed at 1024. We use the Xavier initializer[8] to initialize the model parameters. The optimal hyper-parameters are determined via grid search on the validation set: the learning rate is tuned amongst {0.0001, 0.0005, 0.001, 0.002, 0.005, 0.01}, and finally, we set it to 0.001. The hyperparameter α is searched in $\{10^6, 10^5, 10^4\}$, we set $\alpha = 10^6$. The coefficient β of l_2 regularization is searched in $\{0, 10^5, 10^4, 10^3\}$, we set $\beta = 10^5$. We set the temperature coefficient τ to 0.07. Besides, we stop training if Recall@20 on the validation set does not increase for 10 successive epochs to avoid overfitting.

5.2 Overall Performance

Table 2 illustrates the recommendation performance of the proposed M3CSR and other baselines. According to the table 2, Siamese performs poorly on all datasets, even worse than VBPR which simply

Table 2: Performance comparison of baselines on different datasets in terms of Recall@20, Precision@20 and NDCG@20. The best performance is highlighted in bold and the second to best is highlighted by underlines. Improv. indicates relative improvements over the best baseline in percentage.

Model	Amazon-Sports			Amazon-Baby			Tiktok			Allrecipes		
	R@20	P@20	N@20	R@20	P@20	N@20	R@20	P@20	N@20	R@20	P@20	N@20
VBPR [10]	0.0629	0.0033	0.0277	0.0513	0.0027	0.0227	0.0455	0.0023	0.0185	0.0191	0.0009	0.0074
MMGCN [35]	0.0698	0.0040	0.0279	0.0612	0.0031	0.0241	0.0847	0.0042	0.0268	0.0337	0.0017	0.0123
LATTICE [37]	0.0945	0.0050	0.0444	0.0829	0.0043	0.0371	0.0876	0.0044	0.0380	0.0260	0.0013	0.0097
Siamese [23]	0.0328	0.0017	0.0143	0.0293	0.0015	0.0117	0.0470	0.0023	0.0193	0.0113	0.0008	0.0078
MMSSL [33]	0.0984	0.0052	0.0458	0.0929	0.0049	0.0413	0.0894	0.0045	0.0389	0.0327	0.0016	0.0125
M3CSR (Ours)	**0.1044**	**0.0056**	**0.0477**	**0.0999**	**0.0055**	**0.0436**	**0.1010**	**0.0050**	**0.0419**	**0.0387**	**0.0020**	**0.0143**
%Improv.	6.10%	7.69%	4.15%	7.53%	12.24%	5.57%	12.98%	11.11%	7.71%	14.80%	17.60%	14.40%

Table 3: Ablation study on key components of M3CSR.

Dataset	Amazon-Baby		Tiktok		Allrecipes	
Metrics	R@20	N@20	R@20	N@20	R@20	N@20
w/o MIIL	0.0756	0.0277	0.0911	0.0349	0.0190	0.0129
w/o $\mathcal{L}_{\perp}$	0.0924	0.0425	0.0917	0.0386	0.0328	0.0140
w/o Cluster	0.0775	0.0281	0.0907	0.0378	0.0213	0.0138
w/o Gate	0.0792	0.0295	0.0845	0.0313	0.0257	0.0137
M3CSR	**0.0999**	**0.0436**	**0.1010**	**0.0419**	**0.0387**	**0.0143**

combines multi-modal features using MF. This indicates that solely relying on multi-modal features to address cold-start problems is unrealistic. As the most classic baseline, VBPR only uses multi-modal features as part of the item representation, and LATTICE is more fine-grained to inject higher-level information through the structure between item-item in different modalities. However, they do not deeply model user preferences for content. MMGCN further learns user modality-specific interests on user-item graphs across different modalities, but integrates user multi-modal interests in a unified manner. MMSSL stands as the most competitive baseline, showcasing superior performance among all baselines. It learns the dependencies between collaborative views and multi-modal semantic views while utilizing cross-modal contrastive learning to jointly retain semantic commonalities across modalities. However, we have other more in-depth thinking in multi-modal feature modeling to achieve better performance. Our proposed M3CSR shows promising performance by consistently outperforming all baselines on different datasets, we attribute the performance improvement to leveraging the Modal Encoder to elegantly narrow the gap between the multi-modal embedding extraction and user interest modeling as well as explicit modeling of user multi-modal interest intensity.

5.3 Ablation Study

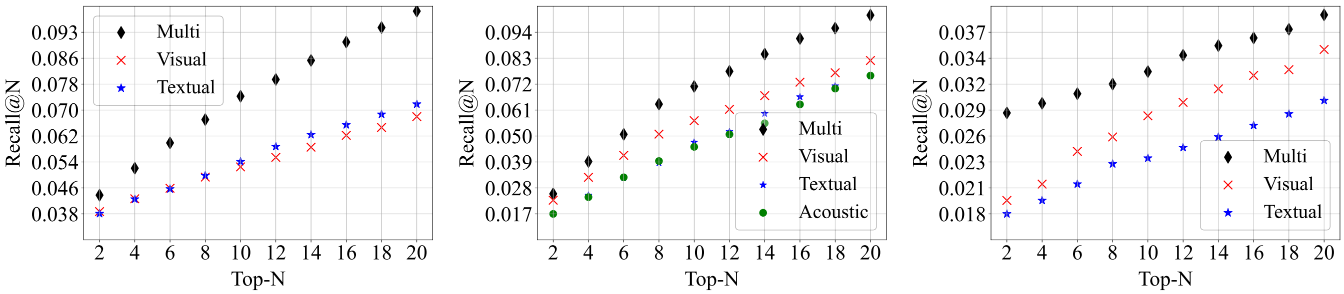
To evaluate the effectiveness of each component in our method, we perform the ablation studies in Table 3. We can see that:

- Without the Multi-modal Interest Intensity Learning network (w/o MIIL), the performance decreases sharply compared with our M3CSR, particularly in the Amazon Baby and Allrecipes datasets. The performance drop is less pronounced in the TikTok dataset, which we attribute to its lower data sparsity and the diverse nature of user modality interests.
- We ablate the pairwise loss with the variant w/o $\mathcal{L}_{\perp}$. Although the experimental results indicate that the gains from decoupling user interests across modalities are minimal on three datasets, it further underscores the importance of maintaining orthogonality in user multi-modal interests learning. Conversely, w/o $\mathcal{L}_{\perp}$ exhibits the most significant performance drop in the TikTok dataset, likely due to its greater diversity in modalities, necessitating stronger decoupling.
- We make another comparison between M3CSR and the variant (w/o Cluster) without the trainable cluster ID. The observed performance gain reflects the improvements of our designed clustering preprocessing in enhancing learnability and addressing the gap between multi-modal feature extraction and modeling.
- Compared with w/o Gate, which removes the gating network of ID embedding and content embeddings on the item-side. We use the interaction frequency of short videos as an indicator of their hotness or coldness, which is based on the total interaction counts across the entire dataset. Experimental results demonstrate the superiority of controlling ID and content embeddings based on the hotness or coldness status of short videos.

5.4 Effects of Modalities

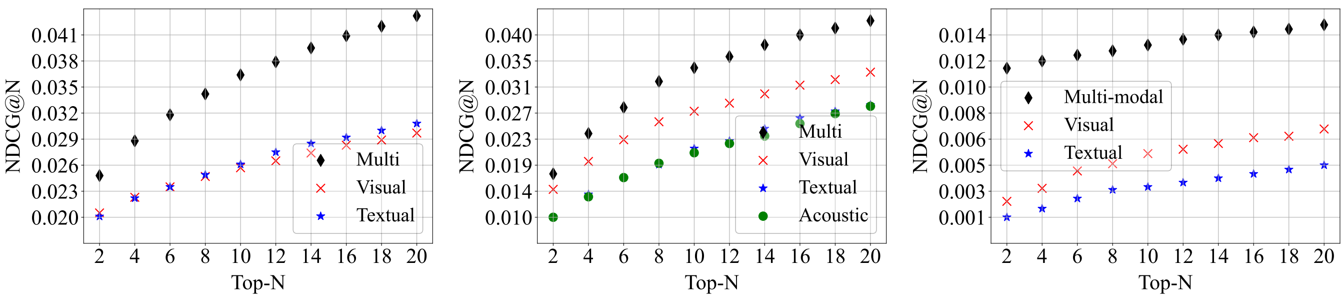
To explore the effects of different modalities, we compare the results on different modalities over the three datasets, as shown in Figure 4. It shows the performance of top-N recommendation lists where N ranges from 1 to 20. We have the following observations:

- As expected, the method with multi-modal features outperforms those with single-modal features in M3CSR on all three datasets. It demonstrates that representing users with multi-modal information achieves higher performance and user preferences are closely related to the content of short-videos. Moreover, it shows that our model could capture the user modality-specific preference from content information.
- On the TikTok and Allrecipes datasets, the visual modality is the most effective among the three modalities. It makes sense because when users browse short videos or food items, people



(a) Amazon Baby (b) Tiktok (c) Allrecipes

Figure 4: Performance in terms of Recall@N w.r.t. different modalities on the three datasets.

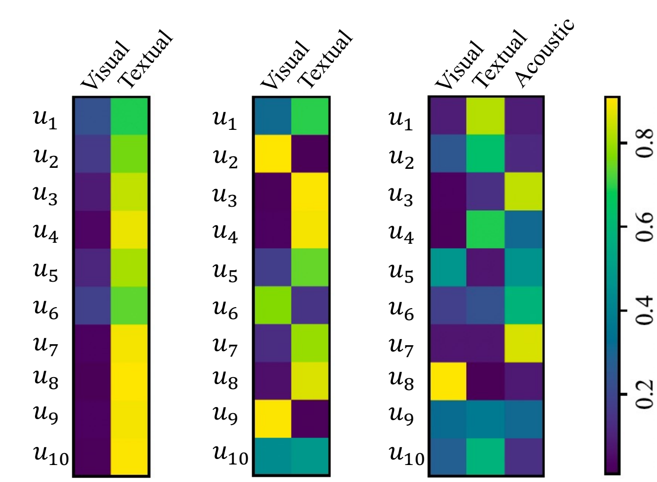


(a) Amazon Baby (b) Tiktok (c) Allrecipes

Figure 5: Performance in terms of NDCG@N w.r.t. different modalities on the three datasets.

usually pay more attention to the visual information than other modality information. However, on the Amazon Baby dataset, the visual modality does not show a substantial advantage, likely due to users emphasizing the product itself when purchasing baby products and not being easily misled by vision.

- The textual modality exhibits the poorest descriptive capability for interaction prediction, particularly on the TikTok and Allrecipes. This is reasonable since we find the text descriptions are of low quality, and even irrelevant to the short-video content on these two datasets. However, this modality offers important cues on the Amazon Baby dataset. Textual descriptions directly represent the functionalities and quality of baby products, being highly relevant to the content. Some users may base their purchases on the functional aspects of baby products. This phenomenon is consistent with our argument that user preference are closely related to the content information.
- Since our dataset only includes acoustic modality information from the TikTok dataset, we observe that acoustic information even has comparable expressiveness to that of the textual modality. It provides valuable insights from another perspective, contributing important information for recommendations.
- We observe that the gap between single-modal modeling and multi-modal modeling is larger on the Amazon Baby and Allrecipes datasets, but smaller on the TikTok. We attribute this to the relatively lower sparsity of the TikTok dataset, which contains richer interaction information. In datasets with higher sparsity, multi-modal joint modeling tends to yield greater benefits.



(a) Amazon Baby (b) Allrecipes (c) Tiktok

Figure 6: Visualization of learned multi-modal interest intensity weights of users selected from three datasets.

5.5 Visualization of Interest Intensity

To further validate the motivation that users exhibit varying intensities of preference across different modalities, we randomly select 10 users from each of the three datasets and show their interest intensity for different modalities in Figure 6. In the Amazon Baby dataset, users predominantly favor the text modality, aligning with

(a) The short videos a user has previously watched. (b) The recommended results provided by M3CSR for this user.

Figure 7: A case study of a user on a real-world short-video platform.

Table 4: The results of A/B testing in online scenario.

Click	Like	Follow	WatchTime	Climbing_{4k}	Coverage
+3.385%	+2.973%	+3.070%	+2.867%	+1.207%	+3.634%

the speculation that users rely more on textual descriptions than visual cues for baby products. On the Allrecipes dataset, users either highly prioritize visual modality or text modality, potentially due to some users being drawn by food covers while others heavily value food reviews. Moreover, in the Tiktok dataset, there is less discrimination in the intensity across modalities compared to Baby or Allrecipes datasets. The reason is that TikTok has more dense interaction data, so the difference in user preferences for different modalities is likely to be less obvious.

5.6 Online A/B Testing

We carried out rigorous online A/B testing in our short-video streaming scenario from Oct. 26, 2023, to Nov. 01, 2023, with hundreds of millions of users per day. In our application, a new or cold-start video is defined as a short video released in less than 24 hours (inclusive) and viewed less than 4,000 times. The results of the online A/B test are shown in Table 4, we focus on several commercial metrics such as *click*, *like*, *follow*, *watch time*, *Climbing* and *Coverage*. *Climbing* is defined as the number of videos whose exposure increases from 0 to a certain count within two days, likening the concept to climbing a peak. For instance, $Climbing_{4k}$ represents the number of videos whose exposure grows from 0 to 4,000 within two days. *Coverage* refers to the proportion of cold-start videos within the recommended results. For company privacy, we cannot report the implementation details and the real performance of the original online models. Instead, we report the performance gain ratio improved by our approach M3CSR. It is worth noting that one percent improvement ratio usually indicates a large improvement of the recommendation capacity in real-world application scenario, when tested on a large population of users. The remarkable online improvements demonstrate the effectiveness of our proposed M3CSR in cold-start recommendation tasks and significantly aided in uncovering high-quality cold-start videos. Our M3CSR is cold-start friendly and avoids excessive bias towards popular videos. Moreover, the growth in the $Climbing_{4k}$ metric indicates that our design in multi-modal modeling has significantly aided in uncovering high-quality cold-start videos.

5.7 Case Study

We show a case study to visually demonstrate M3CSR's precise grasp of user multi-modal preferences. In Figure 7 (a), we present some short-video historical browsing records for a specific user, while (b) displays the recommended results for him/her. Arrows of different colors represent the correlation of different modalities. This user is interested in basketball, fitness, and badminton, our model can recommend cold-start short videos with similar cover images. Even without cover image similarity, we can recommend food- and fitness-related short videos based on textual descriptions that align with the user's interests. Moreover, we can recommend video clips from movies or TV shows with the same soundtrack. One more thing, this user has a greater intensity for visual modality according to the comprehensive recommended results.

6 CONCLUSION

In this paper, we propose a novel and practical multi-modal modeling framework named M3CSR for cold-start short-video Recommendation. We preprocess non-trainable multi-modal features of items to obtain trainable cluster ID, and develop the Modal Encoder to alleviate the gap between the multi-modal embedding extraction and user interest modeling. Additionally, we measure the modality-specific intensity and employ pairwise loss to further decouple the user multi-modal interests. Extensive offline experiments and online A/B testing further verify its effectiveness in cold-start short-video recommendation tasks.

REFERENCES

[1] Yi Cao, Sihao Hu, Yu Gong, Zhao Li, Yazheng Yang, Qingwen Liu, and Shouling Ji. 2022. Gift: Graph-guided feature transfer for cold-start video click-through rate prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2964–2973.

[2] Gaode Chen, Xinghua Zhang, Yijun Su, Yantong Lai, Ji Xiang, Junbo Zhang, and Yu Zheng. 2023. Win-win: a privacy-preserving federated framework for dual-target cross-domain recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4149–4156.

[3] Gaode Chen, Xinghua Zhang, Yanyan Zhao, Cong Xue, and Ji Xiang. 2021. Exploring Periodicity and Interactivity in Multi-Interest Framework for Sequential Recommendation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. 1426–1433.

[4] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 335–344.

[5] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

[6] Manqing Dong, Feng Yuan, Lina Yao, Xiwei Xu, and Liming Zhu. 2020. Mamo: Memory-augmented meta-optimization for cold-start recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 688–697.

[7] Xiaoyan Gao, Fuli Feng, Xiangnan He, Heyan Huang, Xinyu Guan, Chong Feng, Zhaoyan Ming, and Tat-Seng Chua. 2019. Hierarchical attention network for visually-aware food recommendation. *IEEE Transactions on Multimedia* 22, 6 (2019), 1647–1659.

[8] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 249–256.

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

[10] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[12] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. 131–135.

[13] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18, 7 (2006), 1527–1554.

[14] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2333–2338.

[15] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[16] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. 2003. The global k-means clustering algorithm. *Pattern recognition* 36, 2 (2003), 451–461.

[19] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan Kankanhalli. 2019. User diverse preference modeling by multimodal attentive metric learning. In *Proceedings of the 27th ACM international conference on multimedia*. 1526–1534.

[20] 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)*. 94–99.

[21] Qiang Liu, Shu Wu, and Liang Wang. 2017. Deepstyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. 841–844.

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

[23] Sriram Pingali, Prabir Mondal, Daipayan Chakder, Sriparna Saha, and Angshuman Ghosh. 2022. Towards developing a multi-modal video recommendation system. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992.

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

[26] Martin Saveski and Amin Mantrach. 2014. Item cold-start recommendations: learning local collective embeddings. In *Proceedings of the 8th ACM Conference on Recommender systems*. 89–96.

[27] Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, et al. 2021. One model to serve all: Star topology adaptive recommender for multi-domain ctr prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4104–4113.

[28] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1405–1414.

[29] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. 2020. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management* 57, 5 (2020), 102277.

[30] Poonam B Thorat, Rajeshwari M Goudar, and Sunita Barve. 2015. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications* 110, 4 (2015), 31–36.

[31] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[33] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023*. 790–800.

[34] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*. 3541–3549.

[35] 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.

[36] Tianzi Zang, Yanmin Zhu, Haobing Liu, Ruohan Zhang, and Jiadi Yu. 2022. A survey on cross-domain recommendation: taxonomies, methods, and future directions. *ACM Transactions on Information Systems* 41, 2 (2022), 1–39.

[37] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.

[38] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)* 52, 1 (2019), 1–38.

[39] Zhihui Zhou, Lilin Zhang, and Ning Yang. 2023. Contrastive Collaborative Filtering for Cold-Start Item Recommendation. In *Proceedings of the ACM Web Conference 2023*. 928–937.