# Zero-shot Recommendation: Towards Class Semantic Relation Learning for Inferring Labels of Unseen Micro-videos

**Junyang Chen[1,4], Huan Wang[2], Yirui Wu[*3], Qiuzhen Lin[1], Yunfeng Diao[4], Junkai Ji[5]**

[1]College of Computer Science and Software Engineering, Shenzhen University, China
[2]College of Informatics, Huazhong Agricultural University, China
[3]College of Computer and Information, Hohai University, China
[4]Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei University of Technology, China
[5]School of Artificial Intelligence, Shenzhen University, China
junyangchen@szu.edu.cn

## Abstract

Micro-video label prediction plays a pivotal role on contemporary video-sharing platforms, such as Kwai and Tiktok. The emergence of video content lacking labels presents a formidable challenge for conventional user interest prediction methods. This paper addresses the challenge of micro-video label prediction, particularly for unseen videos, by proposing a zero-shot method called Class Semantic Relation Learning (CSRL). Unlike traditional user interest prediction models, CSRL leverages the pre-trained Large Language Model (LLM) to enhance prediction accuracy for unlabeled videos. The novelty of CSRL lies in its integration of three key components: a raw feature autoencoder, LLM-enhanced features, and a decomposed graph network. The decomposed graph network is specifically designed to disentangle the relationships between labeled and unlabeled videos, offering a significant improvement over previous methods. By fusing hidden topics with LLM-enhanced text, CSRL effectively handles sparse video features. Experiments on large-scale datasets from the Kwai platform show that CSRL achieves state-of-the-art results, with up to 44.64% improvement in Hit Ratio (HR), highlighting its superiority over existing zero-shot recommendation models in predicting user interests within the user-video network.

## Introduction

Micro-video label prediction has evolved into a central element of personalized recommendations, as highlighted by the work on personalized recommendation (Liu et al. 2020b; Dai et al. 2022). The recent triumphs of micro-video sharing platforms like Kwai and Tiktok have spurred a growing interest in recommendation systems. Micro-video recommendation can be conceptualized as: given a user-item graph where a portion of videos are labeled, the objective is to predict the labels for the remaining unlabeled ones. Subsequently, recommender can recommend the predicted videos that belong to the same or similar classes as the user's historical video interactions.

A fundamental challenge lies in the fact that only a small fraction of videos are labeled (compared to regular videos, the number of micro-videos generated daily is higher, and they contain less information), while an unlimited number of videos remain unlabeled. This complexity poses a significant hurdle for current methods in inferring labels for these unseen videos (Quan et al. 2025). Traditional classification techniques, such as those outlined in (Chen et al. 2021a,b), confront the same issues and are ill-equipped to handle the emergence of new classes on online sharing platforms. The primary reason behind this limitation is the arduous and costly nature of annotating a sufficient number of labeled micro-videos for novel social topics within content platforms. Consequently, it becomes highly valuable to empower models with the capability to classify videos from these "unseen" classes that lack labeled instances. This facilitates video recommendations based on predicted labels.

Some of previous zero-shot label prediction models have emerged for tackling this challenge (Chen et al. 2023; Wang et al. 2021). Generally, these methods typically structure zero-shot learning (ZSL) in a two-stage manner. The initial step involves acquiring high-quality class semantic descriptions (CSD) as supplementary data. Essentially, they replace the one-hot encoding of known labels with semantically dense embeddings derived from external pre-trained models like BERT (Turc et al. 2019). This step aims to transfer supervised knowledge from seen classes to unseen ones. These models need to perform supervised training with labeled seen videos and subsequently engage in zero-shot learning for unlabeled videos without a predetermined scope. Specifically, previous approaches, including NVIGPN (Chen et al. 2023), DATM (Wang, Zhang, and Gong 2025), and DGPN (Wang et al. 2021), essentially aggregate graph information of seen videos, and then align it with the CSD in the semantic space. However, there are two main drawbacks for these approaches:

a. *Raw features associated with micro-videos may be extremely sparse.* Micro-videos are typically much shorter in length, often ranging from a few seconds to a minute, leading to fewer visual, audio, and textual features being captured over time. This limited duration results in fewer frames, scenes, and content overall, thus making the feature set more sparse. In this way, the previous feature encoding will not perform well.

*Corresponding author

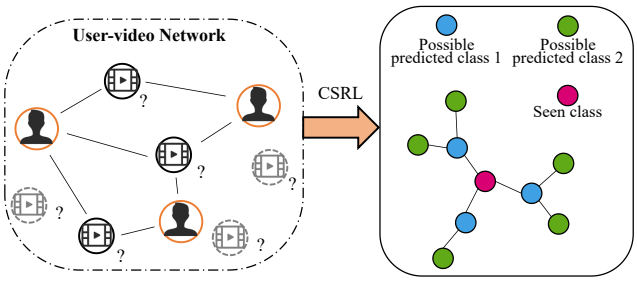



Figure 1: CSRL for zero-shot unseen video label prediction.

b. *Previous methods do not explicitly explore the relationship between labeled videos and unlabeled ones*, i.e., the indirect impact of known labels on the learning of two sets of video labels is unknown. Without considering the relationship between labeled and unlabeled videos, the model may miss out on the latent patterns or correlations between the two sets of videos, which could help in better generalizing to unseen or unlabeled instances. This can lead to suboptimal learning, as the model might not fully leverage the available unlabeled data to improve its performance.

To end with this purpose, we introduce a class semantic relation learning (CSRL) method that is able to exploit the frozen LLM for zero-shot label prediction. As shown in Figure 1, based on users' video browsing history, if the label of a given video is known, CSRL aims to uncover the relationships between videos in order to infer the unseen classes of other videos. First, based on the input user-video data, we leverage raw feature auto-encoder that exploits the hidden topics of video descriptions. Second, to enrich sparse features, we explore LLM to generate short texts of video descriptions. Then, we concatenate the learned topical features and the enhanced features for further relation learning. Third, we explicitly model the seen labeled classes and the unseen ones with a decomposed graph. CSRL can estimate the unseen labeled videos with the learned features and their adjacent relationship. The main contributions include:

- We propose a zero-shot recommendation method that enables label prediction with class semantic relation learning for inferring labels of unseen videos. We exploit the frozen LLM and carefully design three main components in the relation learning.
- We propose to fuse the hidden topics and the LLM-enhanced texts to deal with sparse features associated with videos. We further propose CSRL to explicitly model the seen labeled classes and the unseen ones with decomposed graph network, aiming to disentangle their relationships.
- Extensive experiments on public industrial datasets are conducted to demonstrate that our method outperforms state-of-the-art zero-shot models.

## Related Work

Since limited information contained in micro-videos, we can simply represent them using graph nodes, where node attributes represent the inherent information of the micro-videos. This allows common and efficient graph-based methods to be applied to micro-video label prediction tasks.

### Traditional Graph-based Node Label Prediction

Traditional node label prediction is mainly based on graph neural network (GNN) (Wu et al. 2020), such as GCN (Kipf and Welling 2016), GraphSAGE (Hamilton, Ying, and Leskovec 2017), and GAT (Veličković et al. 2017). These methods have exhibited groundbreaking performance in node label prediction tasks. Subsequently, PinSage (Ying et al. 2018) incorporates GraphSAGE, achieving a significant advancement in deep graph embeddings. This breakthrough opens the door to a new era of large-scale web applications. After that, GNN variants are developed for inductive learning, enabling the inference of unknown nodes and their features. FastGCN (Chen, Ma, and Xiao 2018) interprets graph convolutions as integral transforms of embedding functions and employs Monte Carlo approaches for efficient estimation. DGCN (Zhuang and Ma 2018) proposes a semi-supervised learning method that leverages graph convolution from multiple perspectives to embed graph knowledge, introducing a dual graph convolutional neural network approach to consider both local and global consistency assumptions. HAN (Wang et al. 2019) incorporates both node-level and semantic-level attention to capture the importance of nodes and meta-paths. AdvCaching (Chen et al. 2021b) encourages a target vertex to be close to its neighbors while being far from its negative samples, which is tailored for large-scale network representation learning. Further work, such as VGCL (Yang et al. 2023), leverages GNN to model user-item interactions and employs contrastive learning to enhance node representations. Soft (Zhang and Chen 2024) proposes a soft contrastive framework for sequential recommendation, extending traditional point-to-point contrastive learning to region-level comparisons. DPG (Zheng et al. 2023) constructs a drug interaction graph and employs a GNN to capture relationships between drugs, demonstrating the applicability of traditional graph-based methods in real-world scenarios. In general, these methods can all be applied to micro-video classification. Nonetheless, prevailing methods predominantly assume the existence of labeled nodes for every class within the graph. The incapacity to extend their applicability to unseen classes represents a pivotal challenge confronting current models.

### Zero-shot Graph-based Node Label Prediction

Traditional methods often assume balanced class sizes, which can lead to inaccurate node label prediction when the number of nodes in unseen classes significantly exceeds those in seen classes. Consequently, zero-shot learning (Lampert, Nickisch, and Harmeling 2013) has emerged as a prominent topic, particularly in computer vision, aimed at classifying samples from classes lacking labeled data during training. To address this, class semantic description (CSD) knowledge (Lampert, Nickisch, and Harmeling 2013) facilitates cross-class knowledge transfer, achieved by replacing original one-hot encoded class labels with semantic embeddings derived from external pre-trained models. Subsequently, the predictor learns to recognize these CSDs based on input features during training, enabling inference for unseen class labels by comparing input features with

those of seen classes. This approach effectively aligns label and input feature spaces, constituting a form of transfer learning. Recent graph-based studies have also explored zero-shot learning, such as those proposed by (Wang et al. 2018, 2020), which address imbalanced labels and integrate zero-shot learning into node label prediction. However, these methods often overlook the latent topical information associated with nodes, particularly in video descriptions, where semantic information significantly influences class label prediction. Then, NVGPN (Chen et al. 2023) and DATM (Wang, Zhang, and Gong 2025) consider the importance of the consistency between the external CSD and the input data by exploring the hidden topical information. Moreover, G2P2 (Wen and Fang 2023) leverages graph-based pre-training and prompting techniques to enhance low-resource text classification tasks, including zero-shot and few-shot scenarios. In summary, these methods do not explicitly investigate the correlation between labeled videos and unlabeled ones. In other words, the direct influence of known labels on learning two sets of video labels is not thoroughly examined. Besides, the features associated with items are usually sparse, which makes the zero-shot masks more challenging. Consequently, the predictive performance of existing models on unlabeled videos remains significant room for improvement.

### Distinction and Connection with Traditional Multimedia Short Video Classification

Traditional approaches to short video classification typically rely on modeling three modalities: visual, acoustic, and user viewing history (Lu et al. 2021; Zhang et al. 2016; Xie, Zhu, and Chen 2021), aiming to learn a robust vector representation for each video. For instance, DMLRD (Lu et al. 2021) introduces a low-rank decomposition method to learn low-rank representations of these three modalities while addressing the consistency and complementarity of their features in video representation learning. HMMVED (Xie, Zhu, and Chen 2021) further explores the granularity of visual and acoustic information in short videos, considering the coarse-grained fluctuations in popularity driven by multimodal content. These methods primarily focus on video features, but since short videos are composed of individual frames, the selection of visual and acoustic information from each frame significantly impacts the model's overall performance. To address this, VCE (Zhang et al. 2016) proposes a method that scrapes a set of representative micro-videos from Vine, extracting rich features from textual, visual, and auditory modalities. However, such methods impose significant computational demands, making them unsuitable for the real-time requirements of recommendation systems. As a result, recent approaches simplify the information used from short videos, treating each video as a node where the node's attributes consist of basic information about the video (e.g., likes, views, cover image ID). The primary focus is on the interaction history between users and these nodes, with recommendations being made based on historical interactions and simple labels associated with the micro-videos (Cai et al. 2022). Our method builds upon these simplified approaches and further introduces a technique for classifying short videos under zero-shot conditions.

## Preliminary

### Problem Definition

Let $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{D})$ represent a graph, where $\mathbf{V}$ corresponds to the set of video items $\{v_1, ..., v_{|V|}\}$, $\mathbf{E}$ denotes the edges connecting these items, and $\mathbf{D}$ captures the associated video descriptions. $A \in \mathbb{R}^{|V|\times|V|}$ is an adjacency matrix, where $\mathbf{A}_{ij}$ represents the edge weight between items $v_i$ and $v_j$. Furthermore, we utilize $\mathbf{X} \in \mathbb{R}^{|V|\times T}$ to denote the hidden topical features, signifying that each item is characterized by a $T$-dimensional feature vector. Besides, in this graph, the class set is defined as $\mathcal{C} = \{\mathcal{C}^s \cup \mathcal{C}^u\}$, where $\mathcal{C}^s$ constitutes the seen class set and $\mathcal{C}^u$ comprises the unseen class set, with the constraint that $\{\mathcal{C}^s \cap \mathcal{C}^u\} = \varnothing$. Assuming that all labeled nodes belong to the seen classes, the primary objective of zero-shot micro-video label prediction is to classify the remaining test nodes, characterized by the label set $\mathcal{C}^u$. Additionally, we also explore the relationship between labeled videos and unlabeled ones.

### Hidden Topic Mining

To commence, we employ a neural variational model (Miao, Grefenstette, and Blunsom 2017) to uncover latent topics within the video descriptions. It describes a neural variational model to uncover latent topics from video descriptions. The model extracts hidden topic features by learning a distribution over topics, parameterized by a mean $\mu$ and variance $\sigma$, followed by a re-parameterization trick to generate topic representations. We derive the hidden topic features of the associated video descriptions as follows:

$$\mu = \text{ReLU}(\mathbf{D}\mathbf{W}_\mu), \quad \sigma = \text{ReLU}(\mathbf{D}\mathbf{W}_\sigma), \tag{1}$$

$$\mathbf{X} = \mu + \epsilon \cdot \sigma, \tag{2}$$

where $\mathbf{D} \in \mathcal{R}^{|V|\times|N|}$, $N$ represents the vocabulary set, $\mu$ denotes the mean, $\sigma$ is the variance, $\mathbf{W}_\mu \in \mathcal{R}^{|N|\times T}$ and $\mathbf{W}_\sigma \in \mathcal{R}^{|N|\times T}$ are trainable weights, $T$ is dimension size that we posit the presence of multiple hidden underlying topics, $\epsilon \sim \mathcal{N}(0, 1)$, and $\mathcal{N}$ denotes the standard Gaussian.

### Decomposed Graph Network

To conduct effective graph representation learning, we follow the fundamental ingredients of the variant GCN proposed in (Wang et al. 2021), where the formula is given by:

$$\begin{aligned}\mathbf{V}^{(L)} &= Readout(x) \\ &\approx \theta^* \sum_{i=0}^{L} \binom{L}{i} \beta^{L-i}[(1-\beta)P]^i x,\end{aligned} \tag{3}$$

where $x$ denotes the input matrix of nodes, $\mathbf{V}^{(L)}$ denotes the resulted video representation as the collection of itself and its $\{1, ..., L\}$-hop neighbor information, $\theta^* = \{\theta_1 \cdots \theta_L\}$ denotes the integration of all learnable parameters, $\binom{L}{i}$ is the combinatorial number, and $P = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$, $\tilde{A} = A + I_n$, $I_n$ is identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, $P^i$ represents the matrix $P$ multiplied by itself i times for graph convolution, $\beta$ is ranged in $[0, 1]$ and can be seen as the probability of staying at the current node in a lazy random walk.

## Proposed Model

Our framework in Figure 2 shows the overall architecture consisting of three components: raw feature auto-encoder, LLM-enhanced features, and CSRL with the decomposed graph network.

### Acquiring High-quality CSD for Labels

Initially, we substitute the conventional one-hot encoding of seen labels with semantically rich texts extracted from a large language model (LLM, using GPT-3.5 Turbo). Then, we use TinyBERT (Jiao et al. 2019) to obtain the resulting semantic class description (CSD) embedding, denoted as $\mathbf{s}_c$. In light of the limited availability of labeled data for seen classes, our objective is to leverage the capabilities of the LLM to expand the inference of labels for unseen videos.

### Raw Feature Auto-encoder

We first design an auto-encoder for the raw feature $D$ to obtain its hidden topical features $\mathbf{X}$ by Eq. (2). After that, we propose to minimize the original input $\mathbf{D}$ and the reconstructed output $\hat{\mathbf{D}}$ so as to train the model under unsupervised learning. To end with this, we apply a decoder for $\mathbf{X}$ to obtain the reconstructed input, which is defined by:

$$\hat{\mathbf{D}} = \text{Softmax}(\mathbf{X}\mathbf{W}_d), \tag{4}$$

where $\mathbf{W}_d \in \mathcal{R}^{T\times|N|}$. Then, we define the raw feature loss $\mathcal{L}_{rf}$ with the neural variational topic model (Miao, Grefenstette, and Blunsom 2017) and regularization terms by:

$$\mathcal{L}_{rf} = \sum_{i=0}^{T}(exp(\sigma_i) - (1+\sigma_i) + \mu_i^2), \tag{5}$$

where $\mu_i$ and $\sigma_i$ denotes the mean and variance in each dimension $i$. After the iteration converges (training losses change less than a pre-defined threshold), we adopt the Softmax($\mu$) as the final video-topic distribution to transferable feature learning.

### LLM-enhanced Features

To deal with sparse raw features, we use two stages to enhance them. First, we design a prompt "Generate a definition of no less than 20 words for each phrase" for the associated descriptions of videos and get responses from GPT-3.5. Then, we adopt TinyBert to obtain the enhanced feature embedding, denoted as $\mathbf{E}$. After that, we adopt a feed-forward network (FFN) to the outputs of the concatenation of the above features ($\mathbf{E}$ and $\mathbf{X}$ in Eq. (2)). This FFN layer consists of two linear transformations with a ReLU activation in between as follows:

$$\begin{aligned}\hat{\mathbf{X}} &= \text{FNN}(\mathbf{E}\oplus\mathbf{X}) \\ &= \text{ReLU}((\mathbf{E}\oplus\mathbf{X})\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2,\end{aligned} \tag{6}$$

where $\hat{\mathbf{X}} \in \mathcal{R}^{|V|\times d}$ denotes the learned enhanced features, $\oplus$ denotes the concatenation, $\mathbf{W}_1 \in \mathcal{R}^{(|T|+|N|)\times d}$, $d$ denotes the hidden dimension, $\mathbf{b}_1 \in \mathcal{R}^{|V|\times d}$, $\mathbf{W}_2 \in \mathcal{R}^{d\times d}$, and $\mathbf{b}_2 \in \mathcal{R}^{|V|\times d}$ are trainable parameters.

### Seen Label Inference with DGN

We aim to distill the video relations for transferring the enhanced features $\hat{\mathbf{X}}$ to the label inference. To achieve this purpose, we first construct a video graph for representation learning based on the input user-video data. We connect two videos if they are viewed by the same user. After that, we use decomposed graph network (DGN in Eq. (3)) in the seen label learning. We follow the fundamental ingredients of the DGN used in (Wang et al. 2021). In short, we exploit the combined representations of locality and compositionality in a graph. The adopted locality concerns the small subparts of the input transferable features $\hat{\mathbf{X}}$. We believe that adjacent nodes in the graph structure have closely positioned latent labels in the semantic space. Then, given a semantic embedding $\mathbf{s}_c$ in seen class $\mathcal{C}^s$ and $Readout(\cdot)$ function from Eq. (3), we can have the locality loss $\mathcal{L}_{s\_l}$ as follows:

$$\mathcal{L}_{s\_l} = -\sum_{l=0:L} \ln \frac{\text{sim}(\mathbf{W}_3^T\, Readout(\hat{\mathbf{X}}), \mathbf{s}_c)}{\sum_{c'\in\mathcal{C}^s}\text{sim}(\mathbf{W}_3^T Readout(\hat{\mathbf{X}}), \mathbf{s}_{c'})}, \tag{7}$$

where $sim(\cdot)$ denotes a similarity measure function (here we use the inner product), and $\mathbf{W}_3$ is a trainable matrix. By minimizing the above objective function, each sub-representation of items from 0 to $L$ learns the semantic information from seen classes. After constructing the locality loss, we naturally make the compositionality loss $\mathcal{L}_{s\_c}$ that expresses the learned global representation with a combination of those pre-learned sub-representation:

$$\begin{aligned}&\mathcal{L}_{s\_c} \\ &= -\ln \frac{\text{sim}(\mathbf{W}_4^T \sum_{l=0}^{L} \frac{\binom{L}{l}}{2^L} Readout(\hat{\mathbf{X}}), \mathbf{s}_c)}{\sum_{c'\in\mathcal{C}^s}\text{sim}(\mathbf{W}_4^T \sum_{l=0}^{L} \frac{\binom{L}{l}}{2^L} Readout(\hat{\mathbf{X}}), \mathbf{s}_{c'})},\end{aligned} \tag{8}$$

where $\mathbf{W}_4$ denotes a trainable matrix, $L$ is a pre-defined number of neighbors where we conduct testing in the experimental section, and $\frac{\binom{L}{l}}{2^L}$ denotes scalar weight parameters for normalization (Andreas 2019).

**Seen Label Loss**. We jointly optimize the seen label loss that consists of the compositional loss (Eq. (8)) and the local loss (Eq. (7)) as follows:

$$\mathcal{L}_s = \mathcal{L}_{s\_c} + \lambda\mathcal{L}_{s\_l}, \tag{9}$$

where $\lambda$ is a hyper-parameter to balance the losses, we set it to 1 in the experiment. This jointly learning can simultaneously consider the locality of the node representation and the global compositional representation for the limited labeled node classification.

### Zero-shot Recommendation: Unseen Label Inference with DGN

Given the seen labeled video embeddings $\mathbf{V}_{\mathcal{C}^s}$, we aim to estimate the other unseen labeled ones $\mathbf{V}_{\mathcal{C}^u}$ based on the connectivity. To end with this, we use an intuitive concept, called class semantic relation learning, in the training. We employ Jensen–Shannon divergence $D_{JS}$ to model the unseen label inference. Concretely, we formulate the unseen

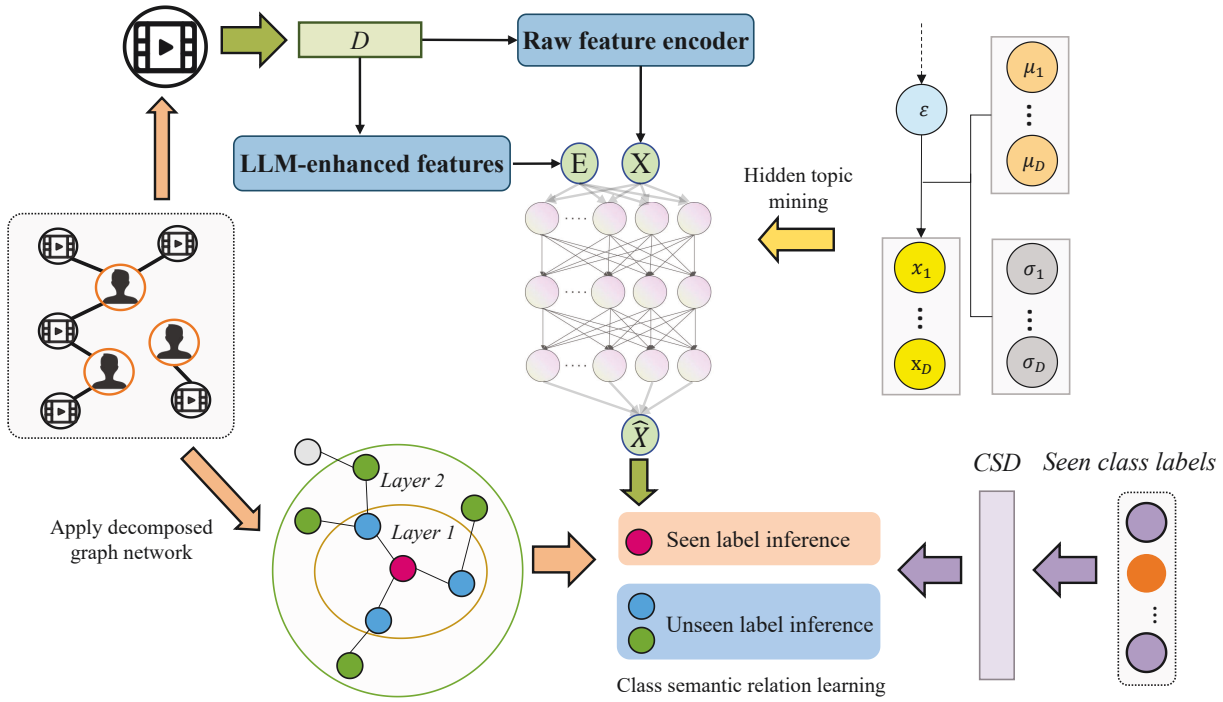



Figure 2: Overview of the proposed CSRL model. It mainly includes four parts: acquiring high-quality CSD, hidden topical mininig with neural variation, seen label learning with graph-based model, and class semantic relation learning.

label loss $\mathcal{L}_{us}$ as follows (Proof is provided):

$$\mathcal{L}_{us} \approx -\ln p(\theta|\mathbf{V}^{(L)}_{\mathcal{C}^u}, \mathbf{s}_c). \tag{10}$$

**Proof:**

$$\begin{aligned}
\mathcal{L}_{us} &= D_{JS}(p(\theta_{\mathbf{us}_c}), \mathbf{s}_c) \\
&= \frac{1}{2} D_{KL}(p(\theta_{\mathbf{us}_c}), p(\theta|\mathbf{V}^{(L)}_{\mathcal{C}^u})) + \frac{1}{2} D_{KL}(p(\theta|\mathbf{V}^{(L)}_{\mathcal{C}^u}), \mathbf{s}_c) \\
&\geq -\ln \sum p(\theta_{\mathbf{us}_c}) \frac{p(\theta|\mathbf{V}^{(L)}_{\mathcal{C}^u})}{p(\theta_{\mathbf{us}_c})} \\
&\quad -\ln \sum p(\theta|\mathbf{V}^{(L)}_{\mathcal{C}^u}) \frac{\mathbf{s}_c}{p(\theta|\mathbf{V}^{(L)}_{\mathcal{C}^u})} \\
&\approx -\ln p(\theta|\mathbf{V}^{(L)}_{\mathcal{C}^u}, \mathbf{s}_c),
\end{aligned} \tag{11}$$

where $\theta_{\mathbf{us}_c}$ denotes the ground-truth distribution of unseen label embedding, $\mathbf{V}^{(L)}$ denotes the video embedding of the final layer from unseen classes, and $\theta$ denotes the estimated distribution of unseen label embedding. We further use a feed-forward layer modeling the semantic relation to estimate $\theta$:

$$\begin{aligned}
&p(\theta|\mathbf{V}^{(L)}_{\mathcal{C}^u}, \mathbf{s}_c) \\
&= \frac{\text{sim}(\text{Softmax}(\tilde{A}^* \mathbf{V}^{(L)}_{\mathcal{C}^s} \mathbf{W}^* + \mathbf{V}^{(L)}_{\mathcal{C}^u}), \mathbf{s}_c)}{\sum_{c' \in \mathcal{C}^s} \text{sim}(\text{Softmax}(\tilde{A}^* \mathbf{V}^{(L)}_{\mathcal{C}^s} \mathbf{W}^* + \mathbf{V}^{(L)}_{\mathcal{C}^u}), \mathbf{s}_{c'})},
\end{aligned} \tag{12}$$

where $\tilde{A}^* = \{A_k \cdots A_1\}$ denotes the k-times multiplication of adjacency matrices, and $\mathbf{W}^* = \{\mathbf{W}_k \cdots \mathbf{W}_1\}$ denotes the trainable matrices (k is set to 3 in the experiments). Note that we assume that the closer two videos are in the adjacency space, the closer they are in the semantic space of labels. The learned $\mathbf{W}^*$ denotes the correlation between labeled videos and unlabeled ones. We believe that by constructing the seen label loss in Eq. (9) and the unseen label loss in Eq. (10), we can effectively disentangle the relationships between labeled (seen) and unlabeled (unseen) videos.

## Joint Learning Loss

Finally, we unify the raw feature autoencoder loss in Eq. (5), the seen label loss in Eq. (9), and the unseen label loss in Eq. (10) as follows:

$$\mathcal{L}_{final} = \mathcal{L}_{rf} + w_1 \mathcal{L}_s + w_2 \mathcal{L}_{us}, \tag{13}$$

where $w_1$ and $w_2$ are set to 1 in the experiments. In general, we optimize $\mathcal{L}_{rf}$ for obtaining the hidden topical features that will be forward into the seen label learning. And $\mathcal{L}_s$ based on the graph learning optimizes the seen labeled video classification with the CSD knowledge. Finally, we estimate the unseen labeled ones with $\mathcal{L}_{us}$.

## Time Complexity Analysis

The computation of our method contains three components: hidden topical mining with neural variation, seen label learning, and class semantic relation learning. As shown in Eq. (13), the first part will cost $O(|V||N|T)$. Then, the learned topical features will be forwarded to graph-based learning. As stated in Section , the second part will cost $O(|V|(|N|+|T|)L|\mathcal{C}^s|d)$. Last, from Section , the third part

costs $O(|V|Kd)$. As a whole, the computational complexity of evaluating Eq. (13) is $O(|V|(|N|T + (|N| + T)L|\mathcal{C}^s| + K)d)$. Detailed explanations of the notions can be referred to the corresponding sections. Moreover, we use sparse-dense matrix multiplications in the implementation. The computational complexity is then $O(|E|(NT+(N+T)L|\mathcal{C}^s|+K)d)$, i.e. linear in the number of graph edges $|E|$.

# Experiment

## Datasets

The detailed statistics of datasets are reported in Table 1. The KuaiRec dataset (Gao et al. 2022) originates from the prominent video-sharing mobile app Kuaishou (accessible at https://kuairec.com/). The dataset is partitioned as follows:

*Class Split I*: there are 7 seen classes further being partitioned to 4 train parts and 3 validation parts. And 23 unseen classes are still used for testing.

*Class Split II*: there are 10 seen classes further being partitioned to 7 train parts and 3 validation parts. And 20 unseen classes are still used for testing.

We evaluate our model on two tasks: zero-shot class label prediction and user interest prediction. For zero-shot prediction, we compare the similarity between the predicted label embedding and class vectors, using the top-k predicted labels for evaluation. For user interest prediction, we use Hit Ratio (HR) and Mean Reciprocal Rank (MRR) to assess recommendation performance based on unseen video interactions. *Details on dataset and experimental settings are provided in the supplementary material.*

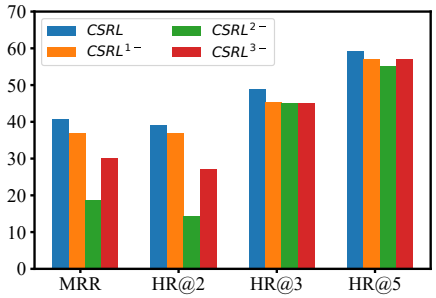
## Comparison Methods

We include several baselines and state-of-the-art models in the comparison: *DAP* (Lampert, Nickisch, and Harmeling 2013), *ESZSL* (Romera-Paredes and Torr 2015), *ZSGCN* (Wang, Ye, and Gupta 2018), *WDVSc* (Wan et al. 2019), *Hyperbolic-ZSL* (Liu et al. 2020a), *DGPN* (Wang et al. 2021), *NVIGPN* (Chen et al. 2023), *DLRSF* (Fan et al. 2024), and *DATM* (Wang, Zhang, and Gong 2025). *The detailed descriptions of the baseline methods are provided in the supplementary material.*

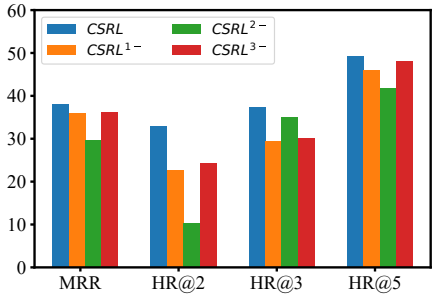
## Zero-shot Recommendation Evaluation from User Interest Prediction Results

Table 2 and Table 3 present the zero-shot user interest prediction results for various methods in both Class Split I and II. In these tables, boldface scores indicate the best results, while underlined scores represent the second-best performance. From these tables, we can observe that:

(1) Table 2 shows that our method, CSRL, consistently achieves the best performance across all metrics in both KuaiRec_Dense and KuaiRec_Sparse datasets. Compared to strong baselines like NVIGPN and DATM, CSRL improves MRR by 0.69% and 0.86%, respectively, while also achieving the highest hit rates (e.g., HR@5 of 38.48% and 42.34%). These results demonstrate that CSRL provides more accurate user interest predictions, benefiting from its ability to integrate LLM-based knowledge and handle unseen classes more effectively.




(a) $KuaiRec\_Dense$.




(b) $KuaiRec\_Sparse$.

Figure 3: Ablation study on two datasets.

(2) Table 3 reports the results for Class Split II. Our method, CSRL, achieves the best performance on both datasets, with an MRR of 40.54% on KuaiRec_Dense and 38.04% on KuaiRec_Sparse. It also leads across all hit rate metrics, e.g., HR@5 reaches 59.15% and 49.32%, respectively. Compared to the strongest baseline, CSRL shows notable gains in MRR and HR@2, demonstrating its strong capability in zero-shot user interest prediction.

## Ablation Study

To demonstrate the effectiveness of each component, we present three variants of CSRL as follows: $CSRL^{1-}$ is a variant model where the raw feature auto-encoder loss (Eq. (5)) is removed; $CSRL^{2-}$ is a variant model where the seen label loss (Eq. (9)) is removed; $CSRL^{3-}$ is a variant model where the unseen label loss (Eq. (10)) is removed. From the comparison results in Figure 3, we observe that:

(1) On KuaiRec_Dense, the performance of our models follows the order: $CSRL > CSRL^{1-} > CSRL^{3-} > CSRL^{2-}$, with CSRL achieving the best results (e.g., MRR 43.11%, HR@5 47.41%). The results suggest that learning from seen labels plays the most critical role, validating our design of separating seen and unseen label learning.

(2) On KuaiRec_Sparse, a similar trend is observed ($CSRL > CSRL^{3-} > CSRL^{1-} > CSRL^{2-}$), with CSRL again achieving the highest performance (e.g., MRR 28.94%). Among ablations, $CSRL^{2-}$ and $CSRL^{3-}$ cause the largest performance drops on all datasets, indicating the importance of decomposed graph and semantic relation loss.

## Investigation the Improvement from LLM

Table 4 compares the performance of CSRL using raw features versus LLM-enhanced features. We observe con-

| Datasets | #Users | #Videos | #User-video | #Video-video | Density | #Classes | Class Split I Train/Val/Test | Class Split II Train/Val/Test |
|---|---|---|---|---|---|---|---|---|
| $KuaiRec\_Dense$ | 1,411 | 3,327 | 4,676,617 | 4,673,748 | 99.6% | 30 | [4/3/23] | [7/3/20] |
| $KuaiRec\_Sparse$ | 7,176 | 10,728 | 10,301,304 | 12,521,016 | 16.3% | 30 | [4/3/23] | [7/3/20] |

Table 1: Statistics of real-world datasets.

| Class Split I | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | KuaiRec_Dense | | | | KuaiRec_Sparse | | | |
| Method | MRR | HR@2 | HR@3 | HR@5 | MRR | HR@2 | HR@3 | HR@5 |
| DAP (Lampert, Nickisch, and Harmeling 2013) | 10.40 | 0.37 | 7.48 | 16.03 | 8.54 | 0.10 | 2.59 | 13.57 |
| Hyperbolic-ZSL (Liu et al. 2020a) | 9.59 | 2.50 | 2.75 | 9.74 | 9.65 | 1.99 | 5.66 | 10.92 |
| WDVSc (Wan et al. 2019) | 16.30 | 9.86 | 14.03 | 22.57 | 17.67 | 11.30 | 15.10 | 23.45 |
| ZS-GCN (Wang, Ye, and Gupta 2018) | 16.39 | 8.18 | 15.28 | 26.05 | 12.75 | 5.97 | 10.51 | 17.83 |
| ESZSL (Romera-Paredes and Torr 2015) | 19.35 | 6.99 | 7.01 | 7.83 | 26.71 | 7.35 | 7.38 | 7.55 |
| DGPN (Wang et al. 2021) | 24.75 | 22.21 | 24.51 | 34.44 | 25.05 | 20.41 | 26.86 | 35.40 |
| NVIGPN (Chen et al. 2023) | 25.92 | 24.41 | 24.98 | 36.72 | 25.88 | 20.27 | 28.19 | 40.84 |
| DLRSF (Fan et al. 2024) | 18.64 | 17.13 | 19.25 | 29.38 | 18.69 | 16.78 | 20.27 | 32.11 |
| DATM (Wang, Zhang, and Gong 2025) | 25.77 | 23.49 | 24.18 | 35.51 | 24.79 | 19.55 | 24.63 | 38.74 |
| CSRL (ours) | **26.10** | **25.71** | **26.84** | **38.48** | **26.94** | **22.43** | **30.24** | **42.34** |
| Improv. | 0.69% | 5.33% | 7.45% | 4.79% | 0.86% | 9.90% | 7.27% | 3.67% |

Table 2: User interest prediction results (%) in Class Split I.

| Class Split II | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | KuaiRec_Dense | | | | KuaiRec_Sparse | | | |
| Method | MRR | HR@2 | HR@3 | HR@5 | MRR | HR@2 | HR@3 | HR@5 |
| DAP (Lampert, Nickisch, and Harmeling 2013) | 13.37 | 7.33 | 11.11 | 20.07 | 13.37 | 7.35 | 11.09 | 21.17 |
| Hyperbolic-ZSL (Liu et al. 2020a) | 9.10 | 2.25 | 4.78 | 5.91 | 9.25 | 3.95 | 7.35 | 9.34 |
| WDVSc (Wan et al. 2019) | 17.70 | 10.65 | 14.86 | 20.18 | 15.29 | 11.90 | 16.53 | 23.70 |
| ZS-GCN (Wang, Ye, and Gupta 2018) | 15.26 | 8.92 | 12.59 | 18.28 | 13.67 | 7.88 | 11.32 | 16.48 |
| ESZSL (Romera-Paredes and Torr 2015) | 27.51 | 15.19 | 15.33 | 15.61 | 23.65 | 12.15 | 14.77 | 14.23 |
| DGPN (Wang et al. 2021) | 21.29 | 14.86 | 18.42 | 24.37 | 31.99 | 29.02 | 33.18 | 40.04 |
| NVIGPN (Chen et al. 2023) | 32.74 | 27.06 | 43.41 | 55.20 | 34.09 | 29.42 | 35.43 | 47.00 |
| DLRSF (Fan et al. 2024) | 28.05 | 19.97 | 20.43 | 21.18 | 25.66 | 17.25 | 18.12 | 18.75 |
| DATM (Wang, Zhang, and Gong 2025) | 30.15 | 21.45 | 22.84 | 23.91 | 27.44 | 20.37 | 26.78 | 39.57 |
| CSRL (ours) | **40.54** | **39.14** | **48.73** | **59.15** | **38.04** | **32.86** | **37.38** | **49.32** |
| Improv. | 23.82% | 44.64% | 12.26% | 7.16% | 11.59% | 11.69% | 5.50% | 4.94% |

Table 3: User interest prediction results (%) in Class Split II.

| Dataset (Class Split II) | KuaiRec_Dense | | KuaiRec_Sparse | |
|---|---|---|---|---|
| Metric | MRR | Hit@3 | MRR | Hit@3 |
| CSRL - raw features | 34.96 | 46.83 | 35.85 | 36.94 |
| CSRL - LLM-enhanced features | 40.54 | 48.73 | 38.04 | 37.38 |

Table 4: Investigation the improvement brought by LLM.

sistent improvements across both KuaiRec_Dense and KuaiRec_Sparse datasets. In terms of MRR, the scores rise from 34.96% to 40.54% on KuaiRec_Dense, and from 35.85% to 38.04% on KuaiRec_Sparse. Similarly, LLM-enhanced features yield slight gains in Hit3, increasing from 46.83% to 48.73% on KuaiRec_Dense and from 36.94% to 37.38% on KuaiRec_Sparse. These results demonstrate that incorporating LLMs helps capture richer feature representations, contributing to improved recommendation quality.

**Supplementary Materials**. Several additional experiments and analyses are provided in the supplementary material. In particular, we examine the impact of varying the hidden topical dimension and the graph layer depth to explore how model complexity relates to representation capability. We also analyze the joint learning loss weights to better understand the trade-offs between different optimization objectives. These studies demonstrate the model's stability across different parameter settings. We commit to releasing the complete version of the codes at the acceptance stage.

## Conclusion

In this paper, we present a zero-shot recommendation approach (CSRL) tailored for micro-video label prediction in user-video networks. CSRL explicitly models the relationships between seen and unseen video classes through a decomposed graph structure and leverages both raw features and LLM-enhanced text to address the challenge of sparse content. By combining hidden topic representations with LLM-derived semantics, our method improves the generalization ability to unseen classes. Experimental results on large-scale datasets from an industrial platform demonstrate that CSRL achieves consistently strong performance compared to existing zero-shot methods. In future work, we can explore the role of user features to enhance the adaptability of zero-shot recommendation models.

## Acknowledgments

This work was supported by Stable Support Project of Shenzhen (20231120161634002), Shenzhen Science and Technology Program (JCYJ20240813141417023), Natural Science Foundation of Guangdong Province of China (2025A1515010233), Guangdong Provincial Department of Education (2024KTSCX060), Tencent "Rhinoceros Birds" - Scientific Research Foundation for Young Teachers of Shenzhen University, Fundamental Research Funds for the Central Universities of China (PA2025IISL0113), Open Project of State Key Lab. for Novel Software Technology of Nanjing University (KFKT2025B22).

## References

Andreas, J. 2019. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*.

Cai, D.; Qian, S.; Fang, Q.; Hu, J.; Ding, W.; and Xu, C. 2022. Heterogeneous graph contrastive learning network for personalized micro-video recommendation. *IEEE Transactions on Multimedia*, 25: 2761–2773.

Chen, J.; Gong, Z.; Mo, J.; Wang, W.; Wang, C.; Dong, X.; Liu, W.; and Wu, K. 2021a. Self-training enhanced: Network embedding and overlapping community detection with adversarial learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11): 6737–6748.

Chen, J.; Gong, Z.; Wang, W.; Wang, C.; Xu, Z.; Lv, J.; Li, X.; Wu, K.; and Liu, W. 2021b. Adversarial caching training: Unsupervised inductive network representation learning on large-scale graphs. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12): 7079–7090.

Chen, J.; Ma, T.; and Xiao, C. 2018. Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247*.

Chen, J.; Wang, J.; Dai, Z.; Wu, H.; Wang, M.; Zhang, Q.; and Wang, H. 2023. Zero-shot Micro-video Classification with Neural Variational Inference in Graph Prototype Network. In *Proceedings of the 31st ACM International Conference on Multimedia*, 966–974.

Dai, Q.; Wu, X.-M.; Fan, L.; Li, Q.; Liu, H.; Zhang, X.; Wang, D.; Lin, G.; and Yang, K. 2022. Personalized knowledge-aware recommendation with collaborative and attentive graph convolutional networks. *Pattern Recognition*, 128: 108628.

Fan, F.; Su, Y.; Liu, Y.; Jing, P.; and Qu, K. 2024. A deep low-rank semantic factorization method for micro-video multi-label classification. *Multimedia Systems*, 30(4): 236.

Gao, C.; Li, S.; Lei, W.; Chen, J.; Li, B.; Jiang, P.; He, X.; Mao, J.; and Chua, T.-S. 2022. 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.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 1024–1034.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3): 453–465.

Liu, S.; Chen, J.; Pan, L.; Ngo, C.-W.; Chua, T.-S.; and Jiang, Y.-G. 2020a. Hyperbolic Visual Embedding Learning for Zero-Shot Recognition. In *CVPR*, 9273–9281.

Liu, W.; Liu, Q.; Tang, R.; Chen, J.; He, X.; and Heng, P. A. 2020b. Personalized Re-ranking with Item Relationships for E-commerce. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 925–934.

Lu, W.; Li, D.; Nie, L.; Jing, P.; and Su, Y. 2021. Learning dual low-rank representation for multi-label micro-video classification. *IEEE Transactions on Multimedia*, 25: 77–89.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*, 2410–2419. PMLR.

Quan, Z.; Chen, J.; Deguchi, D.; Sun, J.; Zhang, C.; Li, Y.; and Murase, H. 2025. Semantic matters: A constrained approach for zero-shot video action recognition. *Pattern Recognition*, 111402.

Romera-Paredes, B.; and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2152–2161.

Turc, I.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wan, Z.; Chen, D.; Li, Y.; Yan, X.; Zhang, J.; Yu, Y.; and Liao, J. 2019. Transductive zero-shot learning with visual structure constraint. In *NIPS*, 9972–9982.

Wang, J.; Zhang, S.; and Gong, Z. 2025. Zero-shot Micro-video Classification with Dual Alignment Topic Model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019. Heterogeneous graph attention network. In *The World Wide Web Conference*, 2022–2032.

Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 6857–6866.

Wang, Z.; Wang, J.; Guo, Y.; and Gong, Z. 2021. Zero-shot node classification with decomposed graph prototype network. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1769–1779.

Wang, Z.; Ye, X.; Wang, C.; Cui, J.; and Yu, P. S. 2020. Network Embedding with Completely-imbalanced Labels. *TKDE*.

Wang, Z.; Ye, X.; Wang, C.; Wu, Y.; Wang, C.; and Liang, K. 2018. RSDNE: Exploring relaxed similarity and dissimilarity from completely-imbalanced labels for network embedding. In *AAAI*, 475–482.

Wen, Z.; and Fang, Y. 2023. Augmenting low-resource text classification with graph-grounded pre-training and prompting. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 506–516.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE TNNLS*.

Xie, J.; Zhu, Y.; and Chen, Z. 2021. Micro-video popularity prediction via multimodal variational information bottleneck. *IEEE Transactions on Multimedia*, 25: 24–37.

Yang, Y.; Wu, Z.; Wu, L.; Zhang, K.; Hong, R.; Zhang, Z.; Zhou, J.; and Wang, M. 2023. Generative-Contrastive Graph Learning for Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1–10. ACM.

Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; and Leskovec, J. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 974–983.

Zhang, J.; Nie, L.; Wang, X.; He, X.; Huang, X.; and Chua, T. S. 2016. Shorter-is-better: Venue category estimation from micro-video. In *Proceedings of the 24th ACM international conference on Multimedia*, 1415–1424.

Zhang, Z. W. W. Y. L. H. P. J. K. G., Yabin; and Chen, X. 2024. Soft Contrastive Sequential Recommendation. *ACM Transactions on Information Systems (TOIS)*, 42(6): 1–28.

Zheng, Z.; Wang, C.; Xu, T.; Shen, D.; Qin, P.; Zhao, X.; Huai, B.; Wu, X.; and Chen, E. 2023. Interaction-aware Drug Package Recommendation via Policy Gradient. *ACM Transactions on Information Systems (TOIS)*, 41(1): 1–32.

Zhuang, C.; and Ma, Q. 2018. Dual graph convolutional networks for graph-based semi-supervised classification. In *Proceedings of the 2018 World Wide Web Conference*, 499–508.