Retrieval-Augmented Hypergraph for Multimodal Social Media Popularity Prediction

Zhangtao Cheng
Jienan Zhang
Xovee Xu
zhangtao.cheng@outlook.com
eroicazjn@outlook.com
xovee@live.com
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Goce Trajcevski
gocet25@iastate.edu
Iowa State University
Ames, Iowa, U.S.A.

Ting Zhong
Fan Zhou*
zhongting@uestc.edu.cn
fan.zhou@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
Kash Institute of Electronics and Information Industry
Kashgar, Xinjiang, China

ABSTRACT

Accurately predicting the popularity of multimodal user-generated content (UGC) is fundamental for many real-world applications such as online advertising and recommendation. Existing approaches generally focus on limited contextual information within individual UGCs, yet overlook the potential benefit of exploiting meaningful knowledge in relevant UGCs. In this work, we propose RAGTrans, an aspect-aware retrieval-augmented multi-modal hypergraph transformer that retrieves pertinent knowledge from a multi-modal memory bank and enhances UGC representations via neighborhood knowledge aggregation on multi-model hypergraphs. In particular, we initially retrieve relevant multimedia instances from a large corpus of UGCs via the aspect information and construct a knowledge-enhanced hypergraph based on retrieved relevant instances. This allows capturing meaningful contextual information across the data. We then design a novel bootstrapping hypergraph transformer on multimodal hypergraphs to strengthen UGC representations across modalities via customizing a propagation algorithm to effectively diffuse information across nodes and edges. Additionally, we propose a user-aware attention-based fusion module to comprise the enriched UGC representations for popularity prediction. Extensive experiments on real-world social media datasets demonstrate that RAGTrans outperforms state-of-the-art popularity prediction models across settings.

CCS CONCEPTS

• Information systems → Retrieval tasks and goals; **• Social and professional topics →** *User characteristics.*

KEYWORDS

Multimedia popularity, hypergraph, retrieval augmentation.

ACM Reference Format:
Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. Retrieval-Augmented Hypergraph for Multimodal Social Media Popularity Prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3637528.3672041

1 INTRODUCTION

With the proliferation of multimodal user-generated content (UGC) including images, texts and micro-videos on social media platforms like TikTok, Triller and Instagram, predicting UGC popularity has become crucial for many real-world applications. For example, recommender systems could leverage popularity forecasts to formulate effective marketing strategies; governments can utilize popularity prediction to identify potential public opinion crises and address them preemptively, avoiding reputation and economic losses [49]; etc. Predicting multimodal social media popularity offers diverse benefits, from helping users navigate information overload to improving downstream applications such as online advertising [30], social recommender systems [23, 46], and rumor detection [35, 58].

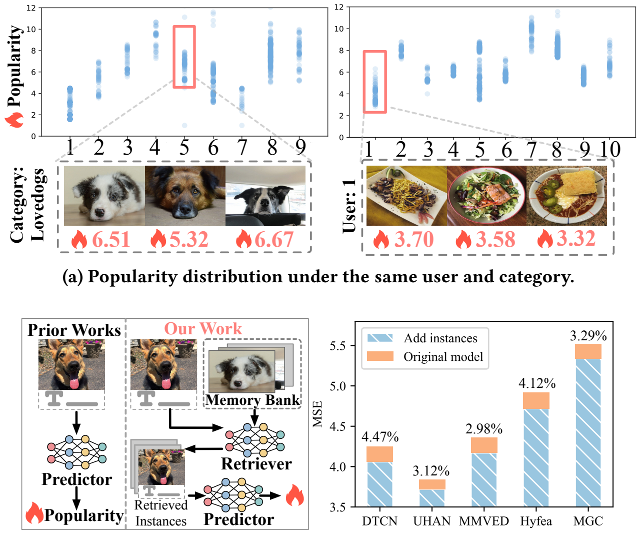
As an important application, multimodal social media popularity prediction (MSMPP) aims to estimate the future popularity of UGC based on multimodal features [48]. Much effort has been devoted to this area, following two main paradigms: (1) *Feature engineering-based methods* prioritize the design of hand-crafted UGC features, including those extracted from user profiles [57], images [27, 37] and time series [40], for directly predicting popularity through well-designed functions. However, a key limitation of these approaches is their dependence on time-consuming manual feature engineering and substantial expert knowledge, constraining their applicability. (2) *Deep learning-based methods* demonstrate promising capabilities in extracting useful patterns directly from UGCs without extensive engineering, achieving substantial performance improvements by learning UGC representations, including both uni-modal [32, 54] as well as more comprehensive multimodal representations [60, 64].

Despite the promising results, prior works explore representation learning only on egocentric UGC, overlooking the benefit of

[*] Corresponding author.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3672041



(a) Popularity distribution under the same user and category.

(b) Prior works *vs.* Ours.

(c) A preliminary experiment.

Figure 1: Motivation of this work. (a) Example of popularity distribution under the same user or category from a Flickr dataset. (b) Prior works mainly use an egocentric UGC for prediction. In contrast, our model retrieves relevant UGCs as an indicator to guide prediction. (c) Comparison between a baseline and the one adding relevant instances. Lower values indicate better performance.

exploiting the useful knowledge from relevant UGCs. However, in real-world social media, follower distributions vary across source users. Thus, even for identical UGCs, social feedback can differ greatly depending on the viewing user group [24]. Fig. 1(a) shows an example of how UGCs posted by different users or in distinct categories tend to exhibit significant differences in popularity distribution. Consequently, modeling just a single UGC in isolation provides limited knowledge, yielding erroneous predictions due to the failure to fully capture diverse popularity distributions.

In contrast, humans have the ability to learn by observation, i.e., the capacity to master skills by observing relevant subjects rather than attempting to memorize every subject [21, 41], which motivates us to improve multimodal social media modeling with retrieved UGCs. As shown in Fig. 1(b), we are specifically interested in designing a retrieval-augmented pipeline. This pipeline searches relevant data instances from the UGC memory bank to generate informative knowledge. The relevant instances act as strong signals and indicators that facilitate popularity prediction. Fig. 1(c) presents a preliminary observation that can further validate our hypothesis. We select 20 data instances from the same user or category in the Flickr dataset, and use the sum of all features as model input. The results suggest that relevant instances can provide valuable information to strengthen model's prediction capability. However, generating expressive knowledge through relevant UGCs for enhancing MSMPP is largely unexplored. Retrieving relevant UGCs from the memory bank and then exploiting meaningful knowledge through relevant instances is non-trivial, primarily due to two major challenges.

C1: Calculating the similarity between the target UGC and relevant instances is complicated, as it requires evaluating multimodal similarities to identify Top-K nearest instances. Existing retrieval methods focus on encoding and retrieving single-modal knowledge [31, 36], and hence cannot make use of multimodal data and their complex correlations. Moreover, UGCs posted on the social platforms typically contain a substantial amount of noise, including discrepancies between textual and visual content, as well as incomplete modal information.

C2: The correlations between the target UGC and retrieved instances are typically high-order. As shown in Fig. 1(a), significant popularity intervals are observed within the same user or category, which suggests a stronger relationship between two distinct UGCs posted even by the same user or belonging to the same category. However, existing methods that directly utilize summation or attention operations [39, 50] to model neighborhood knowledge from relevant instances fail to model the complex high-order correlations between the target UGC and relevant instances.

To address the limitations of prior studies, we present a retrieval-augmented framework for enhancing MSMPP. Our core idea is to leverage the aspect information of UGCs (e.g., category, source user, content topic) and hypergraphs for multimodal retrieval and knowledge augmentation. The aspect information of UGCs helps explicitly discriminate relevant correlations among various UGC types, improving retrieval performance. The underlying intuition is that UGCs with similarities generally exhibit a higher degree of shared aspect information. Besides, a hypergraph can connect more than two vertices via different hyperedges [4], making it well-suited for modeling interrelations among relevant data instances. With this design, hyperedges directly associate multiple vertices by representing UGC aspect information in an explicit manner. Moreover, different hyperedges share common vertices, enabling the modeling of high-order relations between the target UGC and relevant instances through vital shared attributes.

To this end, we propose **RAGTrans**, a novel aspect-aware **R**etrieval **A**ugmented multi-modal hyper**G**raph **Trans**former framework for enhancing MSMPP. Technically, we reformulate the popularity prediction process in a principled *retrieve-and-predict* manner. We construct a UGC memory bank that encodes the visual content, textual content and the aspect information of UGCs with a collection of reference <image, text, aspect> triplets. Specifically, the aspect information involves attribute-level and modal semantic-level information in UGCs, i.e., user, category, topic, textual and visual semantic. During retrieval, we treat each UGC in the memory bank as documents and the target UGC as a query. Retrieving relevant instances could be modeled as a query-document matching problem. We leverage searching techniques and ranking functions (e.g., BM25 [45]) to identify the most informative instances via calculating the similarity scores among their aspect information. During prediction, with the aspect information, we connect all relevant instances via different aspect-level hyperedges and then construct multi-modal hypergraphs to effectively express high-order similar correlations among relevant instances. We also design a bootstrapping hypergraph transformer to extend the information aggregation to the multimodal information mixture, where the intra- and inter-modal propagations are designed to capture the respective intra- and inter-modal correlations. Finally, the model can easily exploit the user-aware cross-attention mechanism to adaptively select the more important and suitable characteristics within UGCs, thus obtaining complementary fused features encompassing multimodal semantics and user-level information for popularity prediction. Following are our main contributions:

- We propose RAGTrans, pioneering an aspect-aware retrieval augmented pipeline that bridges target multimodal UGCs and relevant instances to enhance the MSMPP task.
- We propose a bootstrapping hypergraph transformer that extends information aggregation to the multimodal mixture. Intra-modal and inter-modal propagations are designed to capture correlations within and across modalities as well as fine-grained and aligned UGC representations.
- We conduct extensive experiments on real-world multimodal datasets to evaluate RAGTrans. The results demonstrate that RAGTrans can effectively learn multimodal representations from visual and textual UGC modalities, and achieve up to a 20% gain over strong baseline approaches on the ICIP dataset. The code for reproducing the results is available at https://github.com/CZ-TAO12/RAGTrans.

2 RELATED WORK

2.1 Popularity Prediction.

Predicting the popularity of UGCs in social media is an important problem for various social and recommendation applications [10, 61, 62, 66]. Traditional feature-engineering based methods focus on identifying and incorporating hand-crafted UGC features into machine learning methods. [27] found that image content and social cues (e.g., the number of followers) are beneficial for popularity prediction. In [5], the authors considered both low-level and high-level image features for popularity prediction. Methods like CNN and SVM were used in [6, 29] to extract image latent visual features. [15] analyzed the relationships between visual sentiments and image popularity, and then utilized visual sentiments and user features to predict the popularity scores of social images. Nonetheless, these models rely on the quality of extracted image features and often require extensive domain knowledge. Also, the extracted features are hard to be generalized to new domains.

More recently, deep learning-based methods focus on model architecture designs for modeling different data modalities and fusing them. UHAN [64] designs an user-guided attention network composed of VGGNet and LSTM to merge the visual and textual features for predicting popularity. DTCN [54] integrates ResNet and LSTM to jointly extract neighboring temporal context and periodic temporal context among sequential content. MMVED [60] is based on a variational framework that designs a multimodal variational encoder-decoder for micro-video popularity prediction. MGC [42] and MHF [42] both utilize attention-based mechanisms to learn multimodal UGC data. However, relevant instances contain crucial evidence for guiding the popularity prediction, and existing works have failed to capture meaningful knowledge in relevant UGCs. Our work represents the first step in extending existing methods by retrieving relevant instances to enhance the representations of the target UGC and reason analogously about its popularity score.

2.2 Hypergraph Neural Network.

For graph-structured data in the real-world scenarios (e.g., social networks [9, 13], urban computing [26] and protein structure [59]), graph neural networks express tremendous capacity to model graph-structured data. Graph convolutional network (GCN) [28, 59] is a classical GNN-based framework, which typically performs approximate spectral graph convolution to capture neighborhood structural information of each node in a graph. However, structural interactions in graph-structured data usually exist as many-to-many and high-order relations. Therefore, simple GNNs are incapable of depicting such set-like relations. A *hypergraph* [4] generalizes the edges of a graph into hyperedges, capable to connect an arbitrary number of nodes, which, in turn, provide a natural way to express high-order (set-like) relations. HGNN [14] is the first work to extend graph convolution to hypergraph for capturing the high-order correlations. DHGNN [25] constructs a dynamic hypergraph neural network with two modules: dynamic hypergraph construction and hypergrpah convolution, to model evolution relations among different nodes in the hypergraph. Other variants of hypergraph neural networks (cf. [1]), incorporate diverse hypergraph structures, such as dynamic[25], heterogeneous[47], and motif-based[63] structures.

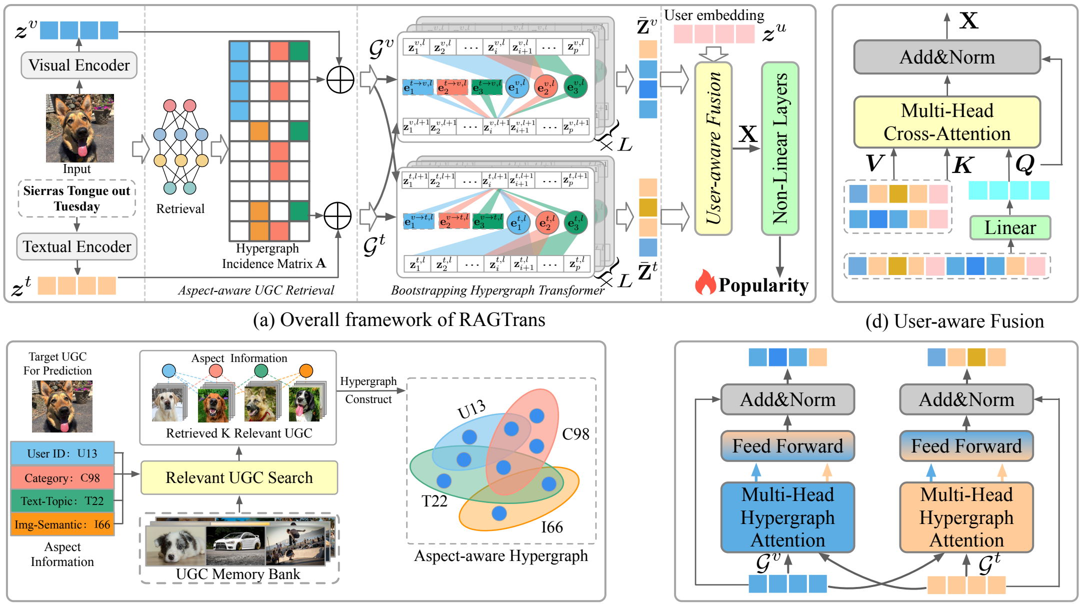
We propose to use hypergraphs to represent complex higher-order correlations between the target UGC and retrieved instances. For each UGC and its surrounding hyperedges, as well as the aspect information from its corresponding node, we obtain expressive representations via our designed hypergraph transformers.

3 METHODOLOGY

As discussed in Sec. 1, using relevant data instances in the memory bank as a kind of auxiliary knowledge is beneficial for accurate prediction of UGC popularity [39]. In this section, we retrieve various relevant data instances according to a *relevance metric* for a given target instance. We aim to use the auxiliary information to help multimodal social media popularity prediction in a unified hypergraph-based framework. After introducing preliminary definitions, we proceed to describe the details of RAGTrans's three main modules. A visual overview of RAGTrans is shown in Fig. 2.

Problem Definition. Let $C = \{c_1, \cdots, c_N\}$ denote the collection of user-generated content (UGC) on social media, where N is number of UGC. For each UGC c_i, we have its multimodal content such as textual descriptions (t) and visual images (v). We aim to learn c_i's multimodal representations $z_i = \{z_i^v, z_i^t, z_i^u\}$, where z_i^v, z_i^t and z_i^u denote visual, textual and user representations, respectively. For the ground-truth popularity y_i for content c_i, it is measured as total interactions of users to c_i in the future, e.g., # reshares, # likes, and # comments. Therefore, given a new content c_i of a user u_i, the task of *multimedia social media popularity prediction* (MSMPP) is to forecast the future popularity y_i via multimodal features z_i.

Hypergraph. Hypergraph is composed of hyperedges that can be used to connect more than two nodes, representing high-order relations among nodes. A hypergraph can be formulated as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the node set and $\mathcal{E}$ is the hyperedge set. The hyperedge $\mathcal{E}$ can be represented using an incidence matrix $\mathbf{A} \in \{0, 1\}^{|\mathcal{V}|\times|\mathcal{E}|}$, where each entry $\mathbf{A}_{v\epsilon}$ is set to 1 if the hyperedge ϵ involves the vertex v, 0 otherwise. In the work, $\mathcal{V}$ is constructed by



(a) Overall framework of RAGTrans

(d) User-aware Fusion

(b) Aspect-aware UGC Retrieval

(c) Bootstrapping Hypergraph Transformer

Figure 2: The structure of the proposed RAGTrans model. (a) The overall framework of RAGTrans, which consists of aspect-aware UGC retrieval, bootstrapping hypergraph transformer, and user-aware fusion. (b) The detailed process of aspect-aware UGC retrieval, introducing search engine techniques to calculate similarity scores. (c) The structure of bootstrapping hypergraph transformer, designed to capture intra- and inter-modal correlations. (d) The detailed structure of user-aware fusion layer.

the nearest instances of the target UGC in the memory bank, and $\mathcal{E}$ denotes the aspect information of UGCs (e.g., user, category).

3.1 Aspect-aware UGC Retrieval

We expect to retrieve relevant data instances of the target UGC and obtain meaningful auxiliary knowledge for guiding the target popularity prediction. First, we construct our UGC memory bank $\mathcal{B}$ that consists of visual content, textual content and the aspect information of massive UGCs with a collection of reference <image, text, aspect> triplets. To use this resource, we design a retrieval module to retrieve Top-K nearest <image, text, aspect> triplets from the memory bank. These retrieved data instances have relations with the target UGC, determined by calculating similarity scores using the aspect information. The aspect information involves different aspects of characteristics, such as attribute-level (i.e., user, category and topic) and modal semantic-level information (i.e., textual and visual semantic), to reflect the finer-grained content in UGCs, enabling a finer-grained assessment of relevance. The detailed construction process of the aspect information is summarized in Appendix A.4. With such designs, leveraging the aspect information to distinguish relevant instances offers a means to mitigate discrepancies between textual and visual content, and enhance retrieval performance. Specifically, the retrieval module is inspired by document retrieval, which allows us to effectively obtain relevant instances from the memory bank via aspect information-based retrieval operations. Following this insight, when there comes a request for predicting c_q, we can abstract data instance c_q into a "document", each aspect information of the instance is regarded as a "term" and then employ search engine techniques [1] to retrieve the nearest data instances in the large memory bank $\mathcal{B}$.

Specifically, we use $\mathbf{x}^q = \{x_f^q | f = 1, \cdots, \mathcal{F}\}$ to represent the feature of aspect information in the target UGC c_q as the query, where x_f^q represents the feature value of f-th field and $\mathcal{F}$ denotes the number of feature fields. First, we implement a boolean query operation to retrieve instances that contain at least one common feature value with the target instance c_q. Next, we use a general search mechanism to calculate relevant score R for each data instance filtered from the whole data space w.r.t. the target content c_q. Then we select K nearest samples from the whole data sample space. Afterward, the ranking function, e.g., BM25 [45], is used to calculate the similarity scores $R(\mathbf{x}^q, \mathbf{x}^D)$ between the query $\mathbf{x}^q$ and a document $\mathbf{x}^D$ in the UGC memory bank:

$$R(\mathbf{x}^q,\mathbf{x}^D)=\sum_{f=1}^{\mathcal{F}} IDF(x_f^q)\frac{TF\left(x_f^q,\mathbf{x}^D\right)\cdot(k_1+1)}{TF\left(x_f^q,\mathbf{x}^D\right)+k_1\cdot(1-b+b\cdot\frac{|\mathbf{x}^D|}{avgdl})}, \quad (1)$$

where k_1 and b are free parameters, and $avgdl$ indicates the average document length. The document length is defined as the number of features in UGC's aspect information, so all documents have the same length $\mathcal{F}$, $\frac{|\mathbf{x}^D|}{avgdl} = 1$. We note $TF(x_f^q, \mathbf{x}^D)$ is the feature x_f^q's term frequency in $\mathbf{x}^D$. If x_f^q is a single value feature, then $TF(x_f^q, \mathbf{x}^D)$ is either 1 or 0 according to whether there is a match of x_f^q in $\mathbf{x}^D$. If x_f^q have multiple feature values, we calculate the

[1] https://github.com/elastic/elasticsearch

Jaccard similarity as the term frequency:

$$TF\left(x_f^q, \mathbf{x}^D\right) = \frac{\left|x_f^q \cap x_f^D\right|}{\left|x_f^q \cup x_f^D\right|}, \quad x_j^D \in \mathbf{x}^D, \tag{2}$$

where $|\cdot|$ denotes the size of a set. IDF is defined as:

$$IDF(x_f^q) = \log \frac{\mathcal{N} - \mathcal{N}(x_f^q) + 0.5}{\mathcal{N}(x_f^q) + 0.5}, \tag{3}$$

where $\mathcal{N}$ is the number of all data samples in the memory bank, and $\mathcal{N}(x_q^t)$ represents the number of filtered data samples that have the same feature value x_f^q in the f-th field. Common features are penalized by the IDF term to be less important than rare features. After the retrieving based on BM25, the Top-K data instances $C_q^{\mathcal{R}} = \{c_q^{r_1}, \cdots, c_q^{r_K}\}$ can be readily obtained. Each retrieved instance $c_q^{r_i}$ contains the <image, text, aspect> triplets.

3.2 Bootstrapping Hypergraph Transformer

Hypergraph [4] can represent cross-instance high-order relations via hyperedges, providing a natural way to connect different relevant instances via aspect-aware hyperedges and distill neighborhood knowledge to assist the inference process on the target UGC. However, UGCs typically consisted of images and texts express users' feelings and opinions. Different modalities will provide diverse information to attract viewers' attention and resulting in different contributions to the popularity of the UGC. It is essential to learn effective and expressive UGC multimodal representations that contain rich intra-modal and inter-modal correlations for assisting inference on the target UGC. However, existing hypergraph-based models, e.g., HGNN [14] and DHGNN [25], are not suitable for multimodal scenarios. They follow a naive strategy for merging isolated uni-modal hypergraphs into a single one through concatenation operations. Since node features propagate on the unimodal hypergraph independently, the learned representations underrate the inherent heterogeneity across multiple data modalities and are insufficient to handle inter-modal correlations.

In order to bridge the isolated unimodal propagations, we design a bootstrapping hypergraph transformer (BHT), to dig into multimodal high-order neighborhood knowledge as well as extracting intra- and inter-relations of multiple data modalities. Intra-relation shows the influence level of neighbors' characteristics on the target instance in the uni-modality, while inter-relation implies that UGC's popularity towards different modalities may correlate. Therefore, BHT extends the node feature aggregation into the multimodal information mixture that contains three stages: intra-modal propagation, inter-modal propagation, and feed-forward network.

Aspect-aware Multimodal Hypergraph Construction. The resulting instances set $C_q^{\mathcal{R}}$ is transformed into a hypergraph $\mathcal{G}_q = \{\mathcal{V}_q, \mathcal{E}_q\}$ of the target instance c_q, where each data instance forms a node, i.e., $\mathcal{V}_q = C_q^{\mathcal{R}}$. Each feature in the aspect information of c_q constructs a hyperedge to connect involved retrieved instances for representing relevant high-order relations between the target UGC and retrieved instances. For example, two UGCs within the same user and category have a stronger relevant correlations than different categories posted by the same user. Moreover, for multimodal content of UGCs, we transform the hypergraph $\mathcal{G}_q$ to one uni-modal hypergraph $\mathcal{G}_q^m$ for each modality $m \in \mathcal{M} = \{t, v\}$, where t and v denote visual and textual modality, respectively. For each node q, it has corresponding uni-modal feature $z_q^m \in \mathbb{R}^d$. The details of modal feature extraction is presented in Appendix A.

Intra-modal Propagation. BHT first propagates on the uni-modal hypergraph to capture intra-modal neighborhood knowledge, which is composed of *node-to-hyperedge* and *hyperedge-to-node* propagations. For the *node-to-hyperedge* propagation, given a query UGC c_q, visual modal representations $\boldsymbol{Z}_q^v = \{z_q^v, z_q^{r_1,v}, \cdots, z_q^{r_K,v}\}$ of hypergraph $\mathcal{G}_q^v$ and hypergraph incidence matrix $\mathbf{A}^v \in \mathbb{R}^{(K+1)\times\mathcal{F}}$, an l-th layer of BHT calculates $\mathcal{F}$ hyperedge representations $\boldsymbol{E}_q^{v,l} = \{\boldsymbol{e}_{q,1}^{v,l}, \cdots, \boldsymbol{e}_{q,\mathcal{F}}^{v,l}\}$ as follows:

$$\boldsymbol{e}_{q,j}^{v,l} = \sum\nolimits_{n_{q,k} \in e_{q,j}^v} \left(\alpha_{jk} \mathbf{W}_z^v \boldsymbol{z}_{q,k}^{v,l-1}\right), \tag{4}$$

where $\mathbf{W}_z^v$ denotes learnable parameters. The superscript l represents the layer of BHT and α_{jk} denotes the attention coefficient of node $n_{q,k}$ in the hyperedge $e_{q,j}^v$. In order to bridge the intra-modal and inter-modal propagation, we inject the information of the inter-modal propagation into the attention coefficient α_{jk}, which can be considered as an influence gate to ensure that only the most influential and informative messages from modality v can be passed through modality t. The attention scores are calculated as follows:

$$\alpha_{jk} = \frac{\exp\left(\sigma\left(\vec{\mathbf{a}}_{az}^T \left[\boldsymbol{z}_{q,k}^{v,l-1} \odot \bar{\boldsymbol{e}}_{q,j}^{v,l}\right]\right)\right)}{\sum_{v_{q,\zeta} \in e_{q,j}} \exp\left(\sigma\left(\vec{\mathbf{a}}_{az}^T \left[\boldsymbol{z}_{q,\zeta}^{v,l-1}, \boldsymbol{z}_{q,\zeta}^{t,l-1}\right] \odot \bar{\boldsymbol{e}}_{q,j}^{v,l}\right)\right)}, \tag{5}$$

where $\vec{\mathbf{a}}_{az} \in \mathbb{R}^d$ is a weight vector, $\bar{\boldsymbol{e}}_{q,j}^{v,l} = \{\boldsymbol{z}_{q,\zeta}^{v,l} | v_{q,\zeta} \in e_{q,j}^v\}$ is the average of the cluster, and $[\boldsymbol{z}_{q,\zeta}^{v,l-1}, \boldsymbol{z}_{q,\zeta}^{t,l-1}] \in \mathbb{R}^{2\times d}$ denotes the concatenation operation. $\odot$ is the Hadamard product. σ is the LeakyReLU activation function.

For the *hyperedge-to-node* propagation, we update presentations $\boldsymbol{Z}_q^{v,l-1}$ via hyperedge features $\boldsymbol{E}_q^{v,l}$:

$$\mathbf{z}_{q,k}^{v,l} = \sum\nolimits_{e_{q,j} \in n_{q,k}^v} \left(\beta_{kj} \mathbf{W}_e^v \boldsymbol{e}_{q,j}^{v,l}\right) \tag{6}$$

$$\beta_{kj} = \frac{\exp\left(\sigma\left(\vec{\mathbf{a}}_{ae}^T \left[\boldsymbol{z}_{q,j}^{v,l-1} \odot \boldsymbol{e}_{q,k}^{v,l}\right]\right)\right)}{\sum_{e_{q,\zeta} \in n_{q,k}^v} \exp\left(\sigma\left(\vec{\mathbf{a}}_{ae}^T \left(\boldsymbol{z}_{q,k}^{v,l-1} \odot \left[\boldsymbol{e}_{q,\zeta}^{v,l}, \boldsymbol{e}_{q,\zeta}^{t,l}\right]\right)\right)\right)},$$

where $\mathbf{z}_{q,k}^{v,l}$ denotes updated representations of node $n_{q,k}^v$ and $\vec{\mathbf{a}}_{ae} \in \mathbb{R}^d$ is a weight vector. β_{kj} represents the attention score of hyperedge $e_{q,j}$ that connects to node $v_{q,k}$.

Inter-modal Propagation. Inspired by the advantages of prefix tuning [34] and corresponding analysis in [19], BHT re-constructs the information propagation process into prefix-guided multimodal information propagation to pre-reduce the modality heterogeneity and captures cross-modal interactions, which distills textual information from $\mathcal{G}_q^t$ to $\mathcal{G}_q^v$. The propagation process from $\mathcal{G}_q^t$ to $\mathcal{G}_q^v$ in the inter-modal propagation can be defined as:

$$\boldsymbol{e}_{q,j}^{t\rightarrow v,l} = \left(\alpha_{jk}^{'} \mathbf{W}_z^t \boldsymbol{z}_{q,k}^{t,l-1}\right), \tag{7}$$

$$\alpha_{jk}^{'} = \frac{\exp\left(\sigma\left(\vec{\mathbf{a}}_{az}^T\left[\boldsymbol{z}_{q,k}^{t,l-1} \odot \bar{\boldsymbol{e}}_{q,j}^{v,l}\right]\right)\right)}{\sum_{v_{q,\zeta}\in e_{q,j}} \exp\left(\sigma\left(\vec{\mathbf{a}}_{az}^T\left[\boldsymbol{z}_{q,\zeta}^{v,l-1}, \boldsymbol{z}_{q,\zeta}^{t,l-1}\right] \odot \bar{\boldsymbol{e}}_{q,j}^{v,l}\right)\right)},$$

where $\alpha_{jk}^{'}$ denotes the attention coefficient of the textual node $n_{q,k}^t$ in the visual hyperedge $e_{q,j}^v$, representing a cross-modal interaction. Especially, above formulas denote the information propagation of textual nodes to visual hyperedges. Then hyperedge-to-node from $\mathcal{G}_q^t$ to $\mathcal{G}_q^v$ can be defined as:

$$\mathbf{z}_{q,k}^{t\rightarrow v,l} = \sum_{e_{q,j}\in n_{q,k}^v} \left(\beta_{kj}^{'} \mathbf{W}_e^t \boldsymbol{e}_{q,j}^{t\rightarrow v,l}\right) \tag{8}$$

$$\beta_{kj}^{'} = \frac{\exp\left(\sigma\left(\vec{\mathbf{a}}_{ae}^T\left[\boldsymbol{z}_{q,j}^{v,l-1} \odot \boldsymbol{e}_{q,k}^{t\rightarrow v,l}\right]\right)\right)}{\sum_{e_{q,\zeta}\in n_{q,k}^v} \exp\left(\sigma\left(\vec{\mathbf{a}}_{ae}^T\left(\boldsymbol{z}_{q,k}^{v,l-1} \odot \left[\boldsymbol{e}_{q,\zeta}^{v,l}, \boldsymbol{e}_{q,\zeta}^{t,l}\right]\right)\right)\right)},$$

where $\beta_{kj}^{'}$ denotes the attention coefficient of the cross-modal interaction. The prefix-guided interaction mechanism prepares a scalar factor to downweight the original attention and redistribute remainder attention for textual modality. Such procedures allow our model to pre-reduce the modality heterogeneity. We also use multi-head attention to expand the model's representation subspaces and stabilize the learning process of self-attention [50]. In this module, we consider endowing our BHT with the capability of jointly attending multi-dimensional dependencies among nodes and hyperedges within aspect-aware hypergraphs. To achieve this, we extend BHT into multi-head hypergraph transformer, which performs head-specific attentive operations in parallel. This extension can be summarized as follows: $Z_{q,h}^{v,l} = \mathrm{HBT}^h\left(Z_q^{v,l-1}, \mathcal{G}_q^v\right), Z_q^{v,l} = \mathrm{Aggregate}\left(Z_{q,h}^{v,l}\right)_{h=1}^{H}$, where Aggregate$(\cdot)$ denotes the concatenation operation. H denotes the number of attention head in the BHT.

Feed-forward Network (FFN). To alleviate modal heterogeneity and obtain fine-grained representations including intra- and inter-modal correlations, we merge the corresponding output features of the prefix-guided interaction module. Inspired by [16] that shows how FFN layer captures task-specific textual patterns, we propose to merge textual hidden states into the visual ones. Given the output of the l-th hyperedge-to-node propagation layer, FFN calculation FNN$(\mathbf{Z}^{v,l})$ is modified as:

$$Z_q^{v,l} = \mathrm{ReLU}\left(Z_q^{v,l}\mathbf{W}_1 + \mathbf{b}_1 + Z_q^{t\rightarrow v,l}\mathbf{W}_3\right)\mathbf{W}_2 + \mathbf{b}_2,$$

$$\bar{\mathrm{Z}}_q^{v,l} = \mathrm{LN}\left(\mathrm{FFN}\left(Z_q^{v,l}\right) + Z_q^{v,l-1}\right). \tag{9}$$

LN [2] is the layer-normalization operation. By merging textual and visual representations into the FFN calculation, RAGTrans aligns the description of the two data modalities. The calculation process of $\bar{\mathrm{Z}}_q^{t,l}$ is same with $\bar{\mathrm{Z}}_q^{v,l}$.

3.3 User-Aware Fusion

Since different modalities have different importance for predicting popularity, and different users have diverse tastes for generating specific textual and visual information in their UGCs (this is even true for the same user), we proposed a user-aware fusion module based on the cross-attention mechanism to automatically select more suitable and important parts in UGC, and in return obtain complementary fused features. Given the output of the final BHT layer – i.e., $\bar{\mathrm{Z}}_q^{t,l}$ and $\bar{\mathrm{Z}}_q^{v,l}$ – we first concatenate two representations from BHT's output with the corresponding user embedding $\mathbf{U} \in \mathbb{R}^{N_u \in d_u}$, and then obtain two new representations: $\boldsymbol{T} = \bar{\mathrm{Z}}_q^{t,l} \| \mathrm{U}$ and $\boldsymbol{V} = \bar{\mathrm{Z}}_q^{v,l} \| \mathrm{U}$. The two representations are modeled to contain high-order neighborhood knowledge, user-lever information, and inter- and intra-modal correlations. They are projected into d dimension as a query. We use both visual and textual representations as key and value and feed them into a user-aware fusion layer:

$$\begin{gathered}\mathrm{ATT}(\boldsymbol{Q}, \boldsymbol{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}, \\ \boldsymbol{X} = \mathrm{ATT}\left(([\boldsymbol{V}\|\boldsymbol{T}]\mathbf{W})\boldsymbol{W}^Q, (\boldsymbol{V}^\top\|\boldsymbol{T}^\top)^\top\boldsymbol{W}^K, (\boldsymbol{V}^\top\|\boldsymbol{T}^\top)^\top\boldsymbol{W}^V\right),\end{gathered} \tag{10}$$

where $\mathbf{W} \in \mathbb{R}^{2(d+d_u)\times(d+d_u)}$ is a learnable parameter matrix, $\boldsymbol{X} \in \mathbb{R}^{d+d_u}$, $\boldsymbol{T}, \boldsymbol{V} \in \mathbb{R}^{2\times(d+d_u)}$ denote output, textual, and visual representations, respectively. $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ denote the query, key, and value projection matrices, respectively. The final popularity is made by fully-connected layers. We use mean squared error (MSE) as the optimization loss.

Table 1: Statistics of three datasets

Dataset	#UGC	# User	#Train	#Val	# Test	V	T
SMPD	305,613	38,312	244,491	30,561	30,561	2,048	384
ICIP	20,337	17,302	16,271	2,033	2,033	2,048	384
WeChat	14,758	18,743	11,808	1,475	1,475	128	128

4 EXPERIMENTS

We now report our experimental results to verify the effectiveness of RAGTrans and address the following research questions:

- **RQ1:** How does our proposed RAGTrans perform on three datasets compared with state-of-the-art baselines?
- **RQ2:** How do different components in RAGTrans contribute to the overall performance?
- **RQ3:** How do the key hyperparameters influence the performance of our RAGTrans?
- **RQ4:** How does the incorporation of retrieval knowledge enhance RAGTrans's prediction performance?

4.1 Experimental Settings

4.1.1 Datasets. We evaluate RAGTrans on three real-world multimedia datasets: SMPD [55], ICIP [38], and WeChat [52]. The detailed statistics of datasets are shown in Table 1. Each dataset is randomly split into training (80%), validation (10%), and test (10%) sets. **V**, **T** denote the dimension of visual and textual features, respectively.

- **SMPD** [55] is a benchmark dataset with multi-faceted information, which is collected from Flickr's online streams and records various social media information from November 2015 to March 2016. SMPD includes user profiles, photo-sharing records, images, and textual descriptions. In this dataset, the popularity is defined as the number of views of each photo.

- **ICIP** [38] dataset records the engagement scores of each Flickr image over a period of 30 days, which contains users' and photos' social features, such as users' groups, image titles, and descriptions. Similarly, the popularity is the number of photo views.
- **WeChat** dataset [52] is collected from the WeChat platform, which tracks WeChat videos over a period of 15 days and records users' interactive behaviors on each video. Furthermore, the dataset contains videos' attributes (e.g., posted time, tags, visual, textual, and acoustic features) and users' interactive characteristics (e.g., stay time, like, comment, click-avatar, and forward). Following previous work [7], the popularity is computed by the mean operation for four popularity indicators: the numbers of comments, likes, reshares, and avatar clicks.

In addition, to suppress the large variations among different UGCs, e.g., view counts of UGC generally vary from zero to millions, we employ a logarithmic function to normalize popularity scores for all datasets [27, 56]: $y = \log_2 r/d + 1$, where y denotes the normalized popularity scores, r is the final view count, d is the number of days since the UGC was posted, and the addition 1 is used to avoid zeros in the logarithm.

4.1.2 Baseline. To evaluate model superiority, we compare RAGTrans with 13 strong baselines, which can be divided into three groups: (1) *Feature engineering*: **SVR** [27] employs support vector regression for popularity prediction, which only considers the visual information of different UGCs. **Hyfea** [29] is a feature-engineering model, which designs hand-craft features and then selects a well-performed model via CatBoost. **MFTM** [22] combines LightGBM and TabNet to capture intricate semantic relationships among different modal content. (2) *Deep learning methods*: **DTCN** [54] considers both neighboring temporal context and periodic temporal context by the temporal attention mechanism. **UHAN** [64] designs a user-guided hierarchical attention to merge both textual and visual features under the guidance of user embeddings. **MMVED** [60] designs a multimodal VAE to encode the input modalities to a low-dimensional stochastic embedding. **MGC** [42] designs a text-guided attention network to learn multimodal UGC data. **MHF** [51] constructs a hierarchical fusion framework to learn multimodal features (i.e., image and text) for image popularity prediction. **CBAN** [11] is to integrate positive attention and negative attention to model relevant and irrelevant information across different modalities. **JAB** [53] assess the impact of a post's title on its popularity while controlling for the time of posting by an attention-based model. **MASSL** [65] constructs a multimodal variational encoder-decoder framework for the popularity prediction. (3) *Hypergraph methods*: **HGNN** [14] introduces graph convolution to hypergraph and designs a hyperedge convolution operation to model the data high-order correlations. **DHGNN** [25] designs a dynamic hypergraph neural network that is composed of two modules: dynamic hypergraph construction and hypergraph convolution.

4.1.3 Implementation Details. All models are tuned to the best performance according to the early stopping strategy when validation errors are not declined for 10 consecutive epochs. For experimental results, we run each model on each dataset five times and report the mean performance. For RAGTrans, model parameters are updated by Adam optimizer and the learning rate is set to 0.001. Furthermore, on SMPD and ICIP datasets, we use pre-trained ResNet to capture the visual features and employ pre-trained sentence-BERT [44] to model the textual features. The dimension of visual and textual features is 2048 and 384, respectively. For the WeChat dataset, we use preprocessed data features. The layer of BHT is 2, and dimension of user embedding is 256. The batch size is 128 and the attention head number H is selected from $\{1, 2, 4, 8, 16\}$. Since these two baselines (HGNN, DHGNN) cannot be directly applied to the multimodal popularity prediction task, we made several adaptions to them. Specifically, model each UGC's category as a hyperedge and each UGC as a node in the hypergraph. For multimodal features, we use a concatenation operation to fuse them.

4.1.4 Evaluation Metric. Following [54, 64], we have chosen two commonly used types of evaluation metrics from the perspective of correlation and precision: Spearman ranking correlation (SRC), mean absolute error (MAE) and mean squared errors (MSE). SRC is used to reflect the ranking correlation between ground-truth popularity y and predicted popularity $\hat{y}$. Higher SRC scores mean better model performance. In addition, we utilize MAE and MSE to compute the average prediction error.

4.2 Overall Performance (RQ1)

The results from comparing RAGTrans with the baselines on the multimodal social media popularity prediction task are reported in Table 2 and we have the following observations:

(O1): Our RAGTrans consistently outperforms all competitive baselines on three datasets under all evaluation metrics. Notably, our model achieves 25.11%, 20.65%, and 22.96% relative gains in terms of MSE, MAE, and SRC on the ICIP dataset, respectively. These experimental results verify the significance of the improvement by RAGTrans. Compared to all baselines, RAGTrans generates more accurate popularity trends, due to its retrieval-augmented strategy. Specifically, RAGTrans effectively retrieves relevant UGC instances and captures expressive knowledge from retrieved instances to guide prediction. Bootstrapping hypergraph transformer facilitates enriched corss-modal high-order relationships between the target UGC and retrieved instances, which further boosts the performance.

(O2): The gaps between feature models and other baselines are relatively small, and in some cases feature models even outperform deep learning models, indicating that deep learning models are not always superior to feature methods. However, its performance heavily relies on hand-crafted features, which are labor intensive and challenging to generalize to new scenarios. This finding is supported by the results of feature models on the SMPD dataset.

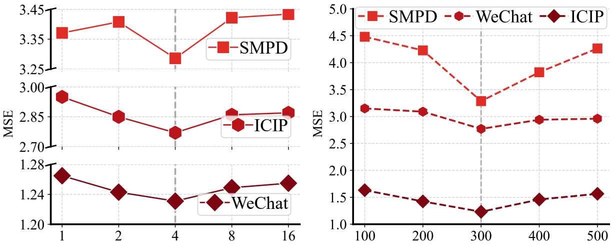
(O3): RAGTrans outperforms deep-learning models by a considerable margin. We attribute the performance deficiency to the inherent incapability of above baselines in effectively exploit crucial clues in pertinent UGCs to guide the target predictions. Moreover, HGNN and DHGNN utilize hypergraph to model cross-modal correlations from multi-modal UGCs, leading to performance improvements. In contrast, RAGTrans decouples the high-order correlations between the target UGC and relevant instances via aspect-aware hyperedges, and captures intra- and inter-modal correlations via the multimodal mixture. The huge performance gap strongly shows the effectiveness of bootstrapping hypergraph transformer in RAGTrans.

Table 2: Performance comparison on three real-world datasets. The best results are in bold font and the second underlined. Lower values of MSE and MAE, and higher values of SRC, indicate better performance.

Dataset	Metric	SVR	Hyfea	MFTM	DTCN	UHAN	MMVED	MGC	MHF	CBAN	JAB	MASSL	HGNN	DHGNN	**RAGTrans**	Improv.
SMPD	MSE	4.9886	4.9297	6.3697	4.2523	3.8471	6.3672	5.5216	3.9297	5.6673	6.1882	13.8925	5.1770	5.0450	**3.2763**	14.83% ↑
	MAE	1.6749	1.6623	1.9590	1.4998	1.4833	1.9607	1.8489	1.5433	1.9058	1.9359	3.1133	1.6061	1.5836	**1.3396**	9.68% ↑
	SRC	0.5312	0.5518	0.3479	0.5432	0.5541	0.2610	0.3228	0.5419	0.1285	0.2353	0.3037	0.4371	0.4698	**0.5859**	5.73% ↑
ICIP	MSE	2.0942	1.9813	1.6268	2.8361	2.7492	1.9831	1.7706	1.8736	3.6143	1.8606	1.8359	1.6711	1.6493	**1.2351**	25.11% ↑
	MAE	1.0552	0.9935	0.8923	1.3432	1.2824	1.0796	1.0117	0.9132	1.3897	0.9289	0.8809	0.9093	0.9010	**0.7149**	20.65% ↑
	SRC	0.3723	0.3641	0.4349	0.3893	0.3981	0.2606	0.3906	0.4041	0.1294	0.3057	0.3937	0.4423	0.4556	**0.5914**	22.96% ↑
WeChat	MSE	2.9551	2.8655	2.8104	3.6921	3.5925	2.9950	2.9450	2.8351	2.9325	2.9654	3.8951	2.9452	2.9031	**2.7928**	1.49% ↑
	MAE	3.2072	3.1073	3.0670	3.4432	3.3132	3.2151	3.1954	3.0543	3.0945	3.1185	3.1294	3.1753	3.1048	**2.9898**	2.11% ↑
	SRC	0.0900	0.1054	0.0794	0.0821	0.0835	0.0911	0.0891	0.1019	0.0706	0.0280	0.0529	0.0939	0.0958	**0.1147**	8.88% ↑

Table 3: Ablation study of RAGTrans.

Module	Variant	SMPD		ICIP	
		MSE	SRC	MSE	SRC
RAGTrans	**All**	**3.2763**	**0.5859**	**1.2351**	**0.5914**
UGC Retrieval	w/o RM	5.5216	0.3228	1.6706	0.3966
BHT Module	w/ $\mathcal{G}_{\mathcal{U}}$	4.9555	0.4083	1.4983	0.4238
	w/ $\mathcal{G}_{C}$	5.0322	0.3745	1.5324	0.4112
	w/o FFN	3.9106	0.4883	1.3787	0.4937
	HyperGAT	4.2481	0.4268	1.4152	0.4587
User-Aware Fusion	w/o **U**	4.9092	0.4143	1.6419	0.4214
	w/o Attn	3.8496	0.5045	1.3211	0.5270



(a) Number of H

(b) Number of K

Figure 3: Sensitivity Analysis of RAGTrans on three datasets. (a) Attention head H in BHT. (b) Retrieved instances K.

4.3 Ablation Study (RQ2)

To gain insights into the major components of RAGTrans, we conduct ablation study on UGC retrieval module, BHT module, and user-aware fusion module. Table 3 shows ablation results.

4.3.1 Effectiveness of UGC retrieval. We first explore the effect of UGC retrieval and verify our motivation for the retrieval-augmented MSMPP. We design a variant model without the UGC *retrieval module* (w/o RM). Based on the results in Table 3, we find that removing the retrieval module would impair prediction performance. It demonstrates that retrieving relevant UGC instances in the memory bank can provide meaningful knowledge for guiding prediction and enhancing the model's generalization.

4.3.2 Effectiveness of BHT module. We explore the effect of hypergraph structures. We compare the performance of the model when we solely use user-level hypergraph (w/ $\mathcal{G}_{\mathcal{U}}$) and solely use the category-level hypergraph (w/ $\mathcal{G}_{C}$). As shown in Table 3, we can observe that using $\mathcal{G}_{\mathcal{U}}$ or $\mathcal{G}_{C}$ severely decrease the prediction performance. This result shows that a large hypergraph containing all UGCs will inject many data noises and lead to suboptimal performance. It further confirms the presence of high-order correlations between the target UGC and relevant instances stored in the memory bank. Moreover, we compare the performance without the feed-forward network (w/o FFN) or with a standard HyperGAT network. The results indicate that both methods harm the prediction performance, and capturing intra- and inter-modal correlations is helpful to obtain expressive UGC representations.

4.3.3 Effectiveness of user-aware fusion module. We build two variants that without user embedding **U** or without fusion attention mechanism. The results indicate that fusion attention mechanism can reduce redundant aggregation of multimodal representations and alleviate negative effects from irrelevant information. In addition, when we remove user embeddings, the model also suffers from a performance drop. This result verifies the effectiveness of our design to inject user information to assist feature fusion.

4.4 Hyper-parameter Analysis (RQ3)

Fig. 3 shows the sensitivity analysis of the RAGTrans parameters.

4.4.1 Multi-head H of BHT. Fig. 3 (a) shows the effect of the multi-head number H of BHT for values $\{1, 2, 4, 8, 16\}$. As can be seen, the performance of RAGTrans initially improves as H increases, and then declines when H is larger, yielding H=4 as appropriate value.

4.4.2 Number of retrieved instances K. Our first observation is that utilizing neighborhood knowledge through the retrieval operation would enhance model performance – in comparison to the variant model *w/o RM*. Retrieval of more data instances boosts the performance of RAGTrans and delivers the best results until $K = 300$. When including more data instances ($K > 300$), the performance of the model starts to decrease. We speculate that this is due to the introduction of noise (irrelevant UGCs) into the model.

4.5 Case Study (RQ4)

4.5.1 Model generalizability. We show the prediction errors of our model, along with the two best-performing baselines (UHAN and HyFea), in Fig. 4. We randomly selected 10 data instances from the SMPD test set, posted by cold start users who have uploaded fewer

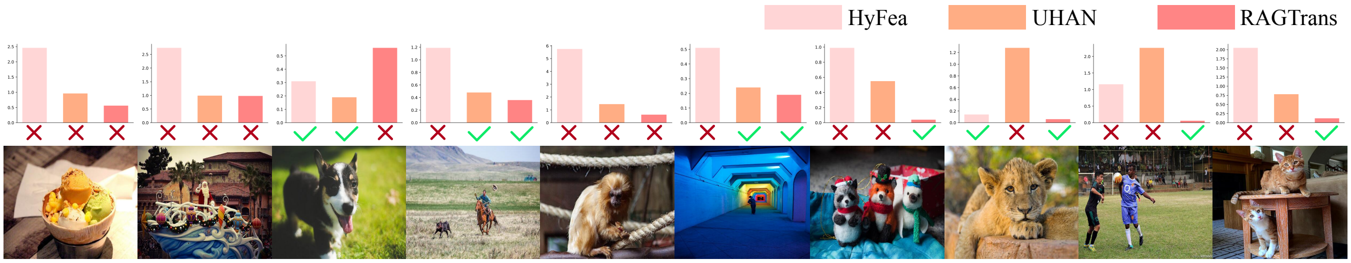


Figure 4: Visualization of model prediction on the SMPD dataset. Top row: error between model prediction scores and ground truth of HyFea, UHAN, and RAGTrans. The higher the bar, the greater the error. Middle row: error threshold (red cross indicates prediction error greater than 0.5; green tick indicates prediction error lower than 0.5. Bottom row: different UGCs.

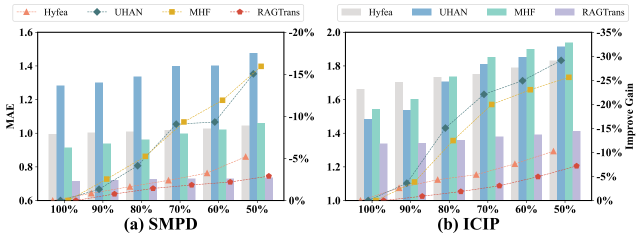
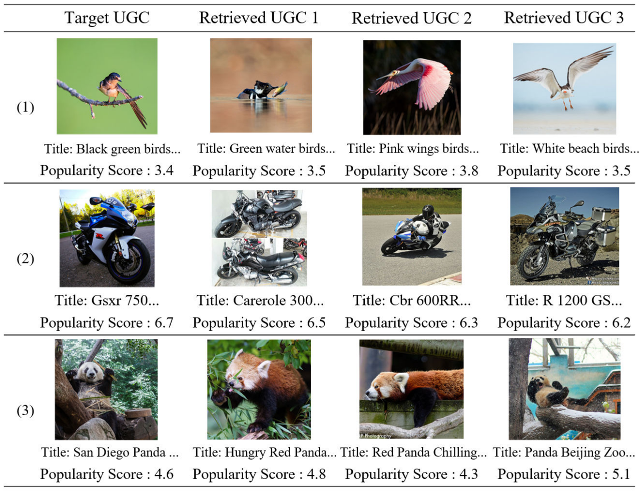


Figure 5: The effect of training set proportion on SMPD and ICIP datasets. The bars represent MAE values and the polylines denote the performance degradation gain.

than three UGCs. We can observe that only two prediction errors of HyFea are lower than 0.5, and only three prediction errors of UHAN are lower than 0.5. For our RAGTrans, there are six UGCs' prediction errors lower than 0.5. Also, RAGTrans exhibits better prediction results on the images that have complex background, which shows that capturing rich and meaningful knowledge from retrieved relevant instances in the memory bank is beneficial for assisting prediction. The results also indicate that employing the retrieval-augmented strategy improves the model generalizability and enhances the prediction performance on newly emerged UGCs.

4.5.2 Model robustness. We investigate the robustness of our model and three best performed baselines (hyfea, UHAN and MHF) by using different portions of the training set. The results are shown in Fig. 5. It is noticeable that the performance of feature engineering method (Hyfea) shows little sensitivity to the size of the training set, while RAGTrans experiences a slight decrease but remains superior to Hyfea. This suggests that Hyfea heavily depends on the quality of hand-crafted UGC features and exhibits limited generalization across different datasets. We can also observe that the performance of deep learning baselines deteriorates significantly when using less training data, as opposed to RAGTrans's slight decrease. This indicates that RAGTrans mitigates the impact of dataset size by capturing meaningful knowledge from retrieved instances and improving UGC representations through the multimodal hypergraph aggregation of intra- and inter-modal neighborhood information. It also shows better robustness of RAGTrans compared to baselines when training data are limited.

4.5.3 Analysis of retrieved instances. We randomly select three UGCs in the test set to visualize top-3 retrieved instances in the memory bank. As shown in Fig. 6, retrieved instances have associated visual content and textual descriptions with the target UGC, verifying the effectiveness of the aspect-aware retrieval.



	Target UGC	Retrieved UGC 1	Retrieved UGC 2	Retrieved UGC 3
(1)	Title: Black green birds... Popularity Score : 3.4	Title: Green water birds... Popularity Score : 3.5	Title: Pink wings birds... Popularity Score : 3.8	Title: White beach birds... Popularity Score : 3.5
(2)	Title: Gsxr 750... Popularity Score : 6.7	Title: Carerole 300... Popularity Score : 6.5	Title: Cbr 600RR... Popularity Score : 6.3	Title: R 1200 GS... Popularity Score : 6.2
(3)	Title: San Diego Panda ... Popularity Score : 4.6	Title: Hungry Red Panda... Popularity Score : 4.8	Title: Red Panda Chilling... Popularity Score : 4.3	Title: Panda Beijing Zoo... Popularity Score : 5.1

Figure 6: Examples of retrieved Top-3 nearest UGCs.

5 CONCLUSION

In this work, we take a pioneering step to design a retrieval-augmented strategy for enhancing UGCs and proposed RAGTrans – an aspect-aware retrieval augmented multi-modal hypergraph transformer to reformulate the prediction process in a retrieve-and-predict format. First, we design a aspect-aware retrieval module to obtain relevant instances. Second, we extend hypergraph attention network into bootstrapping hypergraph transformer to jointly mine intra- and inter-modal information. Finally, a user-aware fusion module is used to obtain fine-grained aligned representations. Extensive experimental results on three datasets demonstrate the effectiveness of our RAGTrans. As part of the future work, we will attempt to incorporate temporal dynamics in the UGC prediction.

6 ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (Grant No.62176043 and No.62072077) and Kashgar Science and Technology Bureau (Grant No.KS2023025).

REFERENCES

[1] Alessia Antelmi, Gennaro Cordasco, Mirko Polato, Vittorio Scarano, Carmine Spagnuolo, and Dingqi Yang. 2023. A Survey on Hypergraph Representation Learning. *Comput. Surveys* (2023).

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[4] Alain Bretto. 2013. Hypergraph theory. *An introduction. Mathematical Engineering. Cham: Springer* (2013).

[5] Ethem F Can, Hüseyin Oktay, and R Manmatha. 2013. Predicting retweet count using visual cues. In *International Conference on Information and Knowledge Management (CIKM)*. 1481–1484.

[6] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. 2015. Latent factors of visual popularity prediction. In *ACM International Conference on Multimedia (MM)*. 195–202.

[7] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In *ACM International Conference on Multimedia (MM)*. 898–907.

[8] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *International Conference on World Wide Web (WWW)*. 925–936.

[9] Zhangtao Cheng, Joojo Walker, Ting Zhong, and Fan Zhou. 2022. Modeling multi-view interactions with contrastive graph learning for collaborative filtering. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[10] Zhangtao Cheng, Wenxue Ye, Leyuan Liu, Wenxin Tai, and Fan Zhou. 2023. Enhancing Information Diffusion Prediction with Self-Supervised Disentangled User and Cascade Representations. In *International Conference on Information and Knowledge Management (CIKM)*. 3808–3812.

[11] Tsun-hin Cheung and Kin-man Lam. 2022. Crossmodal bipolar attention for multimodal classification on social media. *Neurocomputing* 514 (2022), 1–12.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*. 4171–4186.

[13] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The world wide web conference (WWW)*. 417–426.

[14] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33. 3558–3565.

[15] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. 2015. Image popularity prediction in social media using sentiment and context features. In *ACM International Conference on Multimedia (MM)*. 907–910.

[16] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5484–5495.

[17] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 87–110.

[18] John A Hartigan, Manchek A Wong, et al. 1979. A k-means clustering algorithm. *Applied statistics* 28, 1 (1979), 100–108.

[19] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a Unified View of Parameter-Efficient Transfer Learning. In *International Conference on Learning Representations (ICLR)*.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[21] Nicola J Hodges, A Mark Williams, Spencer J Hayes, and Gavin Breslin. 2007. What is modelled during observational learning? *Journal of sports sciences* 25, 5 (2007), 531–545.

[22] Chih-Chung Hsu, Chia-Ming Lee, Xiu-Yu Hou, and Chi-Han Tsai. 2023. Gradient Boost Tree Network based on Extensive Feature Analysis for Popularity Prediction of Social Posts. In *ACM International Conference on Multimedia (MM)*. 9451–9455.

[23] Zheng Hu, Satoshi Nakagawa, Liang Luo, Yu Gu, and Fuji Ren. 2023. Celebrity-aware Graph Contrastive Learning Framework for Social Recommendation. In *International Conference on Information and Knowledge Management (CIKM)*. 793–802.

[24] Liya Ji, Chan Ho Park, Zhefan Rao, and Qifeng Chen. 2023. Neural Image Popularity Assessment with Retrieval-augmented Transformer. In *ACM International Conference on Multimedia (MM)*. 2427–2436.

[25] Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. 2019. Dynamic Hypergraph Neural Networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 2635–2641.

[26] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. 2023. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* (2023).

[27] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular?. In *The Web Conference (WWW)*. 867–876.

[28] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

[29] Xin Lai, Yihong Zhang, and Wei Zhang. 2020. Hyfea: winning solution to social media popularity prediction for multimedia grand challenge 2020. In *ACM International Conference on Multimedia (MM)*. 4565–4569.

[30] Himabindu Lakkaraju and Jitendra Ajmera. 2011. Attention prediction on social media brand pages. In *International Conference on Information and Knowledge Management (CIKM)*. 2157–2160.

[31] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems (Neurips)* 33 (2020), 9459–9474.

[32] Liuwu Li, Runwei Situ, Junyan Gao, Zhenguo Yang, and Wenyin Liu. 2017. A hybrid model combining convolutional neural network with xgboost for predicting social media popularity. In *ACM International Conference on Multimedia (MM)*. 1912–1917.

[33] Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871* (2023).

[34] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4582–4597.

[35] Leyuan Liu, Junyi Chen, Zhangtao Cheng, Wenxin Tai, and Fan Zhou. 2023. Towards Trustworthy Rumor Detection with Interpretable Graph Structural Learning. In *International Conference on Information and Knowledge Management (CIKM)*. 4089–4093.

[36] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. 2022. Retrieval augmented classification for long-tail visual recognition. In *Computer Vision and Pattern Recognition (CVPR)*. 6959–6969.

[37] Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. 2014. " Nobody comes here anymore, it's too crowded"; Predicting Image Popularity on Flickr. In *ACM International Conference on Multimedia (MM)*. 385–391.

[38] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. 2019. Prediction of social image popularity dynamics. In *Image Analysis and Processing (ICIAP)*. 572–582.

[39] Nicolas Papernot and Patrick McDaniel. 2018. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765* (2018).

[40] Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. 2013. Using early view patterns to predict the popularity of youtube videos. In *Web Search and Data Mining (WSDM)*. 365–374.

[41] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526.

[42] Yang Qian, Wang Xu, Xiao Liu, Haifeng Ling, Yuanchun Jiang, Yidong Chai, and Yezheng Liu. 2022. Popularity prediction for marketer-generated content: A text-guided attention neural network for multi-modal feature fusion. *Information Processing & Management* 59, 4 (2022), 102984.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. 8748–8763.

[44] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*. Association for Computational Linguistics (ACL), 3980–3990.

[45] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[46] Lifeng Sun, Xiaoyan Wang, Zhi Wang, Hong Zhao, and Wenwu Zhu. 2016. Social-aware video recommendation for online social groups. *IEEE Transactions on Multimedia* 19, 3 (2016), 609–618.

[47] Xiangguo Sun, Hongzhi Yin, Bo Liu, Hongxu Chen, Jiuxin Cao, Yingxia Shao, and Nguyen Quoc Viet Hung. 2021. Heterogeneous hypergraph embedding for graph classification. In *ACM International Conference on Web Search and Data Mining (WSDM)*. 725–733.

[48] Gabor Szabo and Bernardo A Huberman. 2010. Predicting the popularity of online content. *Commun. ACM* 53, 8 (2010), 80–88.

[49] Alexandru Tatar, Marcelo Dias De Amorim, Serge Fdida, and Panayotis Antoniadis. 2014. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications* 5, 1 (2014), 1–20.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Conference on Neural Information Processing Systems (NIPS)* 30 (2017).

[51] Jing Wang, Shuo Yang, Hui Zhao, and Yue Yang. 2023. Social media popularity prediction with multimodal hierarchical fusion model. *Computer Speech & Language* 80 (2023), 101490.

[52] WeChat. 2021. 2021 China University Computer Contest—WeChat Big Data Challenge. https://algo.weixin.qq.com/2021/problem-description.

[53] Evan Weissburg, Arya Kumar, and Paramveer S Dhillon. 2022. Judging a book by its cover: Predicting the marginal impact of title on Reddit post popularity. In *AAAI Conference on Web and Social Media*, Vol. 16. 1098–1108.

[54] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, and Tao Mei. 2017. Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 3062–3068.

[55] Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. SMP Challenge: An Overview of Social Media Prediction Challenge 2019. In *ACM International Conference on Multimedia (MM)*.

[56] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. 2016. Time matters: Multi-scale temporalization of social media popularity. In *ACM International Conference on Multimedia (MM)*. 1336–1344.

[57] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 30.

[58] Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. 2021. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 1388–1394.

[59] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)* 32, 1 (2020), 4–24.

[60] Jiayi Xie, Yaochen Zhu, and Zhenzhong Chen. 2021. Micro-video Popularity Prediction via Multimodal Variational Information Bottleneck. *IEEE Transactions on Multimedia* (2021).

[61] Xovee Xu, Fan Zhou, Kunpeng Zhang, and Siyuan Liu. 2022. CCGL: Contrastive cascade graph learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 35, 5 (2022), 4539–4554.

[62] Xovee Xu, Fan Zhou, Kunpeng Zhang, Siyuan Liu, and Goce Trajcevski. 2021. CasFlow: Exploring hierarchical structures and propagation uncertainty for cascade prediction. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 35, 4 (2021), 3484–3499.

[63] Junliang Yu, Hongzhi Yin, Jundong Li, Qinyong Wang, Nguyen Quoc Viet Hung, and Xiangliang Zhang. 2021. Self-supervised multi-channel hypergraph convolutional network for social recommendation. In *The Web Conference*. 413–424.

[64] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *The Web Conference (WWW)*. 1277–1286.

[65] Zhuoran Zhang, Shibiao Xu, Li Guo, and Wenke Lian. 2022. Multi-modal Variational Auto-Encoder Model for Micro-video Popularity Prediction. In *International Conference on Communication and Information Processing*. 9–16.

[66] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.

A FEATURE EXTRACTION

The multimodal user-generated content in this work is composed of visual data (image) and textual data (text descriptions). For visual and textual modalities, we use pre-trained ResNet-50 [20] and Sentence-BERT [44] to extract image and text description representations, respectively. In addition, existing feature-based methods had demonstrated that user features (e.g., age and gender) are predictive for popularity prediction [8]. Therefore, we extract user features to improve prediction performance. Notably, our RAG-Trans can effortlessly adapt to alternative pre-trained visual (i.e., CLIP [43] and VIT [17]) and language models (i.e., Bert [12] and AnglE [33]).

A.1 Visual Modality

For visual knowledge, they can intuitively reflect UGC semantics and attract users' attention. Given a set of images $\mathcal{I} = \{I_1, \ldots, I_N\}$, we use pre-trained ResNet-50 to extract visual representations $Z^v = \{z_1^v, \ldots, z_N^v\} \in \mathbb{R}^{N \times d}$. Specifically, the calculation process can be defined as:

$$z_i^v = \mathrm{ResNet}(I_i). \tag{11}$$

A.2 Textual Modality

For textual knowledge, they provide valuable UGC descriptions along with the visual content. Given a set of UGC descriptions $\mathbf{Te} = \{Te_1, \ldots, Te_N\}$, we employ pre-trained Sentence-BERT [44] to extract textual representations $Z^t = \{z_1^t, \ldots, z_N^t\} \in \mathbb{R}^{N \times d}$ as follows:

$$z_i^t = \text{Sentence-BERT}(Te_i). \tag{12}$$

A.3 User Embedding

To customize users' identities and capture users' divergent influences, we describe a user u with a learnable embedding vector $\mathbf{u} \in \mathbb{R}^{d_u}$, where d_u denotes the adjustable vector dimension. Particularly, all user embedding vectors formulate a user embedding matrix, which can be seen as an embedding look-up table: $\mathbf{U} = \{\mathbf{u}_1 \ldots \mathbf{u}_{N_u}\}$. Here N_u is the number of users. Note that the user embedding matrix $\mathbf{U}$ is regarded as the initial state of users and can be dynamically updated via the label of the studied UGC in an end-to-end fashion.

A.4 Aspect Information Construction

For features $\mathbf{x}^i = \{x_f^i | f = 1, \ldots, \mathcal{F}\}$ in the aspect information of data instance c_i in Section 4.1, we design four types of UGC aspect information (i.e., user ID, category, text topic and image semantic) and employ search engine techniques to retrieve the K nearest data instances of the target. For image semantics, we use clustering operation (i.e., k-means clustering [18]) to partition the visual features Z^v into $\mathcal{K}_v$ clusters in which each observation belongs to the cluster with the nearest mean (cluster centroid), serving as a prototype of the cluster. Specifically, $\mathcal{K}_v$ clusters denote $\mathcal{K}_v$ types of image semantics. For text topics, we employ the topic model Latent Dirichlet allocation (LDA) [3] as the topic clustering tool to obtain $\mathcal{K}_t$ text topics. For the remaining attributes, they are collected from the original data source.