Improving Multimodal Social Media Popularity Prediction via Selective Retrieval Knowledge Augmentation

Xovee Xu, Yifan Zhang, Fan Zhou, Jingkuan Song*

University of Electronic Science and Technology of China, Chengdu, Sichuan 610054 China
xovee.xu@gmail.com, yifanzhang@std.uestc.edu.cn, fan.zhou@uestc.edu.cn, jingkuan.song@gmail.com

Abstract

Understanding and predicting the popularity of online User-Generated Content (UGC) is critical for various social and recommendation systems. Existing efforts have focused on extracting predictive features and using pre-trained deep models to learn and fuse multimodal UGC representations. However, the dissemination of social UGCs is not an isolated process in social network; rather, it is influenced by contextual relevant UGCs and various exogenous factors, including social ties, trends, user interests, and platform algorithms. In this work, we propose a retrieval-based framework to enhance the popularity prediction of multimodal UGCs. Our framework extends beyond a simple semantic retrieval, incorporating a meta retrieval strategy that queries a diverse set of relevant UGCs by considering multimodal content semantics, and metadata from user and post. Moreover, to eliminate irrelevant and noisy UGCs in retrieval, we introduce a new measure called Relative Retrieval Contribution to Prediction (RRCP), which selectively refines the retrieved UGCs. We then aggregate the contextual UGC knowledge using vision-language graph neural networks, and fuse them with an RRCP-Attention-based prediction network. Extensive experiments on three large-scale social media datasets demonstrate significant improvements ranging from 26.68% to 48.19% across all metrics compared to strong baselines.

1 Introduction

Social media popularity prediction (SMPP) is a critical task across many domains, such as social networking services (Hong, Dan, and Davison 2011) and online marketing (Aggrawal et al. 2017; Gu et al. 2024), benefiting applications ranging from information propagation and rumor detection (Moniz and Torgo 2019) to social recommendation and network traffic management (Zhang et al. 2021; Cheng et al. 2022). At its core, SMPP research seeks to understand what factors influence the diffusion of user-generated content (UGC) among users and develop prediction models that can accurately "foresee" the future popularity of UGCs.

Early efforts in SMPP focused on mining predictive patterns from UGC content and social context, exploring various UGC features and building machine learning models for the prediction (Tatar et al. 2014; Liu et al. 2022; Zhou et al. 2021). For example, visual features such as color patches, gradient, and objects in images, alongside social features like follower count and historical UGC popularity, were found to be effective in (Khosla, Das Sarma, and Hamid 2014). Another research direction employed statistical models to simulate UGC diffusion processes (Zhao et al. 2015; Mishra, Rizoiu, and Xie 2016), including survival analysis and point processes. More recent attention has been paid to UGC representation learning through neural networks, which automatically capture complex UGC patterns and integrate multiple data modalities into a unified model, leading to state-of-the-art prediction performance (Cheung and Lam 2022; Cheng et al. 2024; Hsu et al. 2023; Chen et al. 2023; Xu et al. 2023). For example, the vision-language Transformers have been used in (Cheung and Lam 2022) to learn visual and textual representations of UGC content; and the hierarchical variational auto-encoders have been used in (Xie, Zhu, and Chen 2023) to model UGC's internal noises and external uncertainties.

A major challenge in SMPP lies in identifying the intrinsic quality and diffusion patterns of UGCs that significantly impact future popularity. However, many exist-

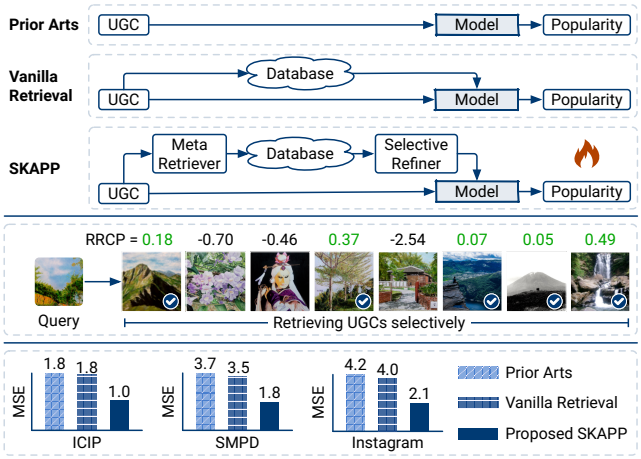

Figure 1: Top: Model comparison between non-retrieval, vanilla retrieval, and our proposed selective retrieval. Middle: Case study of selectively retrieving UGCs from the Instagram dataset. Bottom: Preliminary experiments.

*Corresponding Author

ing approaches treat UGC prediction as an isolated process, focusing heavily on semantic learning while overlooking the interconnected nature of UGCs. These connections, whether explicit or implicit, often arise from users' social ties and shared community interests (Ferrara, Interdonato, and Tagarelli 2014). Furthermore, UGC popularity is not solely driven by social relations, but is equally influenced by social exposures, such as trends, events, and personalized recommendations (Abbar, Castillo, and Sanfilippo 2018).

To capture the social relations and UGC-User interactions, a plethora of methods have designed user and diffusion graphs, leveraging graph neural networks (GNNs) (Cao et al. 2020; Ji et al. 2023b) to learn structural relationships between UGCs or users. However, these efforts are limited to "peaking strategy"-based prediction at where the early diffusion patterns are observed, and require users' social networks that may not always be available or functional for cold-start users. Retrieval augmentation models, as an alternative way of using contextual knowledge to enhance generation and prediction, have attracted a lot of attention across various domains, including large language models (Gao et al. 2023; Long et al. 2024), visual question answering (Lin et al. 2024), and popularity prediction (Ji et al. 2023a; Zhong et al. 2024).

Retrieval-augmented approach meets the goal of modeling UGC relations and interactions, with a focus on UGC itself and its semantic and social contexts. Despite its effectiveness in enhancing contextual learning, we found that a retrieval-based framework for SMPP task encounters the following significant challenges:

- A simple retrieval strategy that relies solely on semantic similarity often fail to reflect the overall contextual information of complex social UGCs. For example, when the query is only concerned with the UGC photo, we retrieve many UGCs having similar photos – but their dynamics can be very different from the query in users and topics – neglecting other potentially relevant UGCs.
- Not all retrieved UGCs may be truly relevant to the query UGC. The quality of the retrieval results heavily depends on the design of the query and the retrieval algorithm, which are not always reliable or optimal (Cuconasu et al. 2024) – as a consequence, inevitably introducing noises and irrelevant UGCs that could be harmful to the prediction. This is especially the case for social UGCs that have varied quality and informal content.

To address these challenges, we present SKAPP, a Selective retrieval Knowledge Augmentation framework for multimodal social media Popularity Prediction. Beyond a simple semantic retrieval strategy, we propose a *meta retriever* that considers not only multimodal UGC semantics, including vision-enhanced language descriptions, but also the social contexts of UGCs by incorporating metadata information, such as user and post dynamics. Inspired by conditional cross-mutual information (Fernandes et al. 2021; Wang et al. 2023), we devise a *selective refiner* based on a new measure termed Relative Retrieval Contribution to Prediction (RRCP). The selective refiner quantifies the gains in prediction of the retrieved UGCs conditioned on the query, filtering out potentially irrelevant and noisy UGCs to the prediction. Moreover, to effectively aggregate multimodal knowledge from the selected UGCs, we introduce vision-language GNNs for UGC contextual learning, coupled with an RRCP-Attention-based prediction network for multimodal knowledge fusion and final UGC popularity prediction. A comparison between traditional model, vanilla retrieval, and our proposed SKAPP is sketched in Figure 1.

Overall, we show that the meta retriever, selective refiner, and prediction network work together to achieve universal performance improvements, with gains ranging from 26.68% to 48.19% across all metrics on three large-scale datasets compared to cutting-edge baselines. Further ablation studies validate the effectiveness and robustness of our proposed SKAPP model. Source codes and datasets are available at https://github.com/YifanZhang-git/SKAPP.

2 Preliminaries

Problem Definition Given a set of N multimodal user-generated content (UGC), denoted as $C = \{c_1, c_2, \ldots, c_N\}$, the problem target is to predict their future popularity $P = \{p_1, p_2, \ldots, p_N\}$, e.g., the number of likes or views. Each UGC c_i is represented by a triplet (c_i^v, c_i^t, c_i^m), where c_i^v corresponds to the visual modality, c_i^t to the textual modality, and c_i^m to the metadata modality. The metadata modality includes user and post information, such as the number of friends, UGC tags, posting time and location.

Retrieval-Augmented Generation Retrieval-augmented generation (RAG) is a technique that combines the strengths of information retrieval systems and generative models (Gao et al. 2023). It has emerged as a way of to enhance the performance of large language models (LLMs) by integrating external knowledge into the response generation. The common flow of RAGs involves first searching and retrieving relevant knowledge from a large database based on the query, followed by fusing the retrieved knowledge with the input to enhance learning and generation. Despite its effectiveness, RAG faces several limitations, including the *irrelevance of retrieved knowledge* and the *noise in retrieval* (Cuconasu et al. 2024; Kevin Wu 2024). Motivated by RAG and also its limitations, we aim to build a retrieval-based SMPP framework that can retrieve the most relevant UGCs while eliminating the irrelevant and noisy UGCs during retrieval.

3 Methodology

Overview SKAPP model consists of three key modules: the meta retriever, selective refiner, and knowledge-augmented prediction network. The meta retriever is designed to identify a diverse set of relevant UGCs by considering multimodal UGC content and metadata. The selective refiner measures the retrieved UGCs based on their relative prediction contributions, aiming to filter out irrelevant and noisy UGCs. The prediction network employs vision-language GNNs and an RRCP-Attention-based module for better fusing the retrieved knowledge. The framework of SKAPP is depicted in Figure 2.

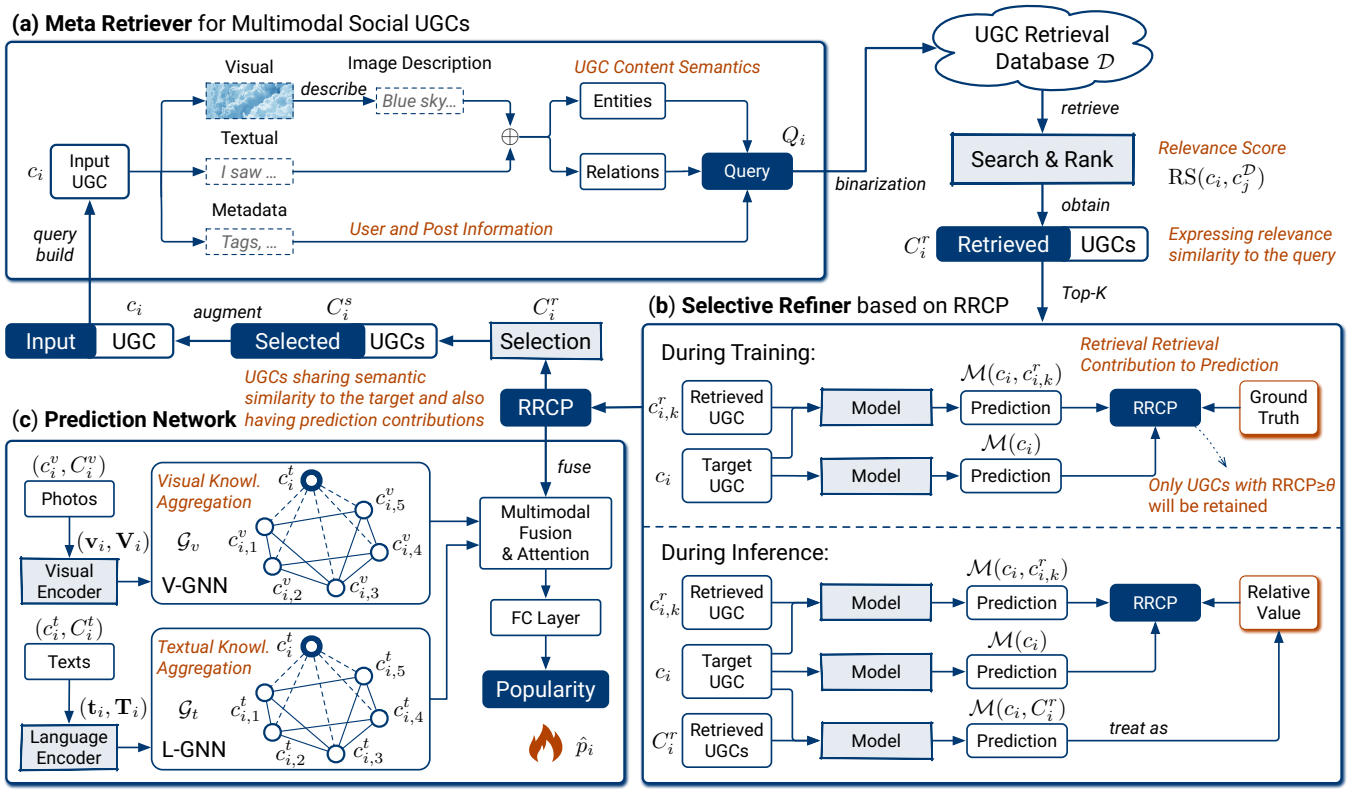

Figure 2: Framework of the proposed SKAPP model. The input is a social UGC and the output is its predicted future popularity. (a) *Meta Retriever* constructs the UGC query by integrating multimodal UGC content semantics with the metadata information, enabling the retrieval of a broader and more diverse set of potential relevant UGCs. (b) *Selective Refiner* employs a new Relative Retrieval Contribution to Prediction (RRCP) measure to select UGCs that have positive gains in prediction, filtering out irrelevant and noisy UGCs. (c) *Prediction Network* leverages vision-language graph neural networks to aggregate contextual knowledge from selected UGCs with an RRCP-Attention-based module for accurate social media popularity prediction.

Meta Retriever for Multimodal Social UGCs

Social UGCs are multimodal content created by social media users, whose behavior differs markedly from that of professionals (Momeni, Cardie, and Diakopoulos 2015). These UGCs often feature informal language, slang, and low-quality visuals, and their dissemination is influenced by user interests, social trends, and platform algorithms. The high diversity and variability of social UGCs pose significant challenges for effective retrieval. Beyond a simple content retrieval strategy, we propose meta retriever, which considers not only the multimodal information, including visual and language semantics, but also the multi-faceted nature of social UGCs by incorporating contextual understanding of the metadata information, such as user, posting time, tags, and friends.

Multimodal UGC Query Construction Given a social UGC c_i as the retrieval query, we expect to retrieve the most relevant UGCs from a knowledge base. The retrieved UGCs should express high semantic relevance to the query and share similarities in user, post, social trend, and event information that can be helpful for the prediction. Relying on unimodal information, such as visuals, may yield UGCs with matching images but divergent in other critical aspects like user demographics or topic relevance – neglecting many other potentially relevant UGCs. The query construction process in meta retriever involves: (i) we first extract visual semantics from UGC's visual modality, c_i^v, using a pre-trained image-to-text model (Li et al. 2022), which generates a rich visual description; (ii) this visual description is integrated with the UGC's textual content, c_i^t; (iii) then entities and actions are extracted from the combined texts, thereby enriching the query with detailed content and contextual information; (iv) the final query, Q_i, comprises these enhanced textual descriptions with metadata such as user Id, tags, and categories, ensuring a comprehensive and contextually aware query for retrieval.

Meta Retrieval With the query Q_i prepared, containing vision, language, and metadata information of UGC c_i, we proceed to retrieve relevant UGCs from a UGC knowledge database $\mathcal{D}$. We use the BM25 ranking function (Robertson et al. 1995), a standard tool in information retrieval, to search through $\mathcal{D}$ and rank UGCs based on their relevance to Q_i. Specifically, we calculate a relevance score, $\mathrm{RS}(c_i, c_j^{\mathcal{D}})$, to assess the similarity between the query UGC c_i and each

UGC $c_j^{\mathcal{D}}$ in the database:

$$\text{RS}(c_i, c_j^{\mathcal{D}}) = \sum_{f \in Q_i} \text{IDF}(c_{i,f}) \cdot \frac{T(c_{i,f}, c_{j,f}^{\mathcal{D}}) \cdot (k_1 + 1)}{T(c_{i,f}, c_{j,f}^{\mathcal{D}}) + k_1} \quad (1)$$

where k_1 is a BM25 hyperparameter, $T(c_{i,f}, c_j^{\mathcal{D}})$ measures the feature similarity between $c_{i,f}$ and $c_{j,f}^{\mathcal{D}}$, and $\text{IDF}(c_{i,f})$ is the inverse frequency of the feature $c_{i,f}$ across UGCs. Since retrieving UGCs differs from traditional document retrieval, the term $(1-b+b\cdot|c_j^{\mathcal{D}}|/\text{avgdl})$ is simplified to 1. For similarity measure $T(c_{i,f}, c_{i,f}^{\mathcal{D}})$, $T\left(c_{i,f}, c_j^{\mathcal{D}}\right)$ equals 1 if $c_{i,f} = c_j^{\mathcal{D}}$, otherwise 0. If $c_{i,f}$ is a vector containing multiple values, we use the Jaccard similarity defined as:

$$T\left(c_{i,f}, c_{j,f}^{\mathcal{D}}\right) = |c_{i,f} \cap c_{j,f}^{\mathcal{D}}| / |c_{i,f} \cup c_{j,f}^{\mathcal{D}}|. \quad (2)$$

We note that the above intersection operation is feasible for small-sized UGC knowledge databases, but it becomes infeasible for large databases due to the increased computation and reduced relevance of retrieved UGCs. To improve the retrieval effectiveness for large databases, we modify the Jaccard similarity to focus only on exact matches rather than intersections, ensuring the most relevant UGCs are prioritized. The database size is determined by whether we could still retrieve many valid UGCs when applying a more strict measure. Our preliminary experiments also verified the efficiency and effectiveness of this approach. Using the calculated relevance scores, the top-K most relevant UGCs are retrieved from the knowledge base $\mathcal{D}$, denoted as $C_i^r = \{c_{i,1}^r, c_{i,2}^r, \cdots, c_{i,K}^r\}$.

Selective Retrieval

Why we need selective retrieval? On the one hand, social UGCs they themselves are inherently containing more noise and irrelevant content than professional, well-organized content. Their topics and diffusion dynamics are complex and multi-faceted, influenced by both user preferences and community trends. The effectiveness of the retrieval is highly dependent on the quality of the queries used. On the other hand, unlike traditional retrieval strategies, our proposed meta retriever expands the retrieving search space – when we retrieve more diverse UGCs that could be useful, we also retrieve UGCs potential of being noisy and irrelevant for the prediction, resulting in suboptimal performance. Therefore, while maintaining the breadth of the meta retriever, we are interested in a way of selecting retrieved UGCs that distinguishes the useful UGCs. From this perspective, the final retrieved UGCs should not only be semantically and metadata-wise similar to the target, but also contribute positively to the prediction.

Inspired by conditional cross-mutual information (CXMI) used in neural machine translation (Fernandes et al. 2021) and large language models (Wang et al. 2023), we propose to measure the prediction contribution of retrieved UGCs, retaining those with high prediction contributions and filtering out others. CXMI is a measure quantifies the influence of context on model's prediction, specifically how much information the context provides about prediction given input data. In our case, this concept can be abstracted to *how much the retrieved UGCs contribute to the prediction*.

Derived from the CXMI, we design a new measure termed Relative Retrieval Contribution to Prediction (RRCP), tailored for social UGCs and the SMPP task. RRCP measures the relative performance change with and without the retrieved contextual UGCs, conditioned on the query UGC. Moreover, we use a directional contribution approach for estimating RRCP during inference. At last, the obtained RRCP scores can also benefit the multimodal UGC fusion network, which we will detail later.

Relative Retrieval Contribution to Prediction Given any arbitrary UGC learning model, say $\mathcal{M}$, we can predict c_i's popularity by $\hat{p}_i = \mathcal{M}(c_i)$, or by augmenting the input with retrieval knowledge: $\hat{p}_i = \mathcal{M}(c_i, c_{i,k}^r)$, $c_{i,k}^r \in C_i^r$. Then RRCP is quantified as the difference in prediction errors:

$$\text{RRCP}(c_i, c_{i,k}^r) = \mathcal{L}\left(p_i, \mathcal{M}\left(c_i\right)\right) - \mathcal{L}(p_i, \mathcal{M}(c_i, c_{i,k}^r)), \quad (3)$$

where p_i is the ground-truth popularity of c_i and $\mathcal{L}(\cdot,\cdot)$ can be any loss function of interest. Here we use the mean absolute error – a positive RRCP value indicates a beneficial contribution from the retrieved UGC $c_{i,k}^r$. Conversely, a negative value indicates that $c_{i,k}^r$ is redundant or introducing noises, and is therefore filtered out. For estimating the RRCP for a held-out test set, a directional contribution approach is used to quantify the relative prediction contribution. It hypothesizes that a trained model can differentially learn from various contextual UGCs and that the cumulative prediction gain from an unfiltered set of retrieved UGCs is greater than a single retrieved UGC. The RRCP for UGC $c_{i,k}^r \in C_i^r$ queried by c_i is estimated by:

$$\mathcal{L}(\mathcal{M}(c_i, C_i^r), \mathcal{M}(c_i)) - \mathcal{L}(\mathcal{M}(c_i, C_i^r), \mathcal{M}(c_i, c_{i,k}^r)). \quad (4)$$

This direction contribution approach is simple and flexible, relieving the burden of training a specialized estimation model with a different prediction target (Wang et al. 2023) (which expressiveness and scalability are limited and we empirically found that it is infeasible for complex social UGCs), or conducting Monte Carlo simulations (Fernandes et al. 2021). After obtaining the RRCP values for each $c_{i,k}^r$ in C_i^r, the selection of the retrieved UGCs is performed as:

$$C_i^s = \{c_{i,k}^r \mid c_{i,k}^r \in C_i^r, \text{RRCP}(c_i, c_{i,k}^r) \geq \theta\}, \quad (5)$$

where θ is the threshold for selection.

Perspective from Recall and Precision The meta retriever and selective refiner we proposed in this work are analogous to the well-established metrics of Recall and Precision, respectively (refer to Figure 3). The meta retriever is designed to retrieve broader and more diverse UGCs that might be missed by simpler retrieval methods – increasing the proportion of relevant UGCs that were retrieved, i.e., optimizing Recall. On the other hand, the selective refiner focuses on refining the pool of retrieved UGCs by their contributions to the prediction, akin to optimizing Precision – increasing the proportion of relevant UGCs among all retrieved ones. This ensures the retrieval process not only captures a wide range of potentially useful UGCs but also maintains high relevance and utility in its outcomes.

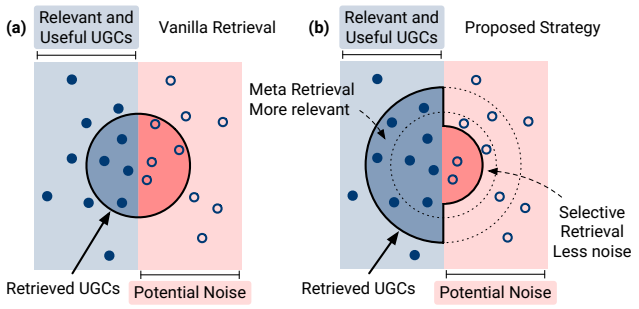

Figure 3: Illustration of meta retriever and selective refiner.

Knowledge-Augmented Prediction Network

Now we have obtained the selected UGCs C_i^s that relevant to the query c_i and have prediction contributions quantified by the RRCP. Here we introduce vision-language GNNs based on GraphAdapter (Li et al. 2024) and an RRCP-Attention-based prediction network. We first model the visual and textual modalities of UGCs via a vision GNN and a language GNN, respectively, treating UGCs as nodes and their relations as edges. By using a graph structure, the contextual knowledge of the query UGC can be effectively aggregated within the visual or language graph. Then a multimodal fusion module based on the RRCP measure attentively aggregates the vision and language knowledge from two graphs. At last, the fused knowledge is fed into fully-connected (FC) layers for the final popularity prediction.

Vision-Language Graph Neural Networks First, two pre-trained vision and language encoders are used to extract each UGC's visual $\mathbf{v}$ and textual $\mathbf{t}$ embeddings. The vision graph is defined as $\mathcal{G}_v = (\mathcal{N}_v, \mathcal{E}_v)$, with nodes $\mathcal{N}_v = \{c_i^v | 1 \leq i \leq |C_i^s| + 1\}$ being the set of all UGC nodes including the query and all selected UGCs, and $\mathcal{E}_v = \{e_{i,j}^v \mid 1 \leq |C_i^s| + 1, 0 \leq j \leq |C_i^s| + 1\}$ is the set of edges. For each node c_i^v, its attribute is the corresponding visual embedding $\mathbf{v}_i$. The edge value is defined as the cosine similarity between the visual embeddings of two nodes: $e_{i,j}^v = (\mathbf{v}_i \cdot \mathbf{v}_j)/(\|\mathbf{v}_i\|\|\mathbf{v}_j\|)$. A similar structure is used for the textual graph $\mathcal{G}_t = (\mathcal{N}_t, \mathcal{E}_t)$. Both graphs reinsert the query node c_i^t to enhance the representation learning of the UGC interactions, with node correlations as edge values. Afterwards, two graph convolution networks are used to obtain visual $\mathbf{Z}_i^v = \text{V-GNN}(c_i, C_i^s)$ and textual $\mathbf{Z}_i^t = \text{T-GNN}(c_i, C_i^s)$ representations of the UGCs.

RRCP-Attention-based Fusion & Prediction Instead of a simple pooling or concatenation, we make use of the previously obtained RRCP values with an attention mechanism to fuse the learned visual and textual representations, applying two levels of weighted knowledge aggregation:

$$\mathbf{z}'_{i,v} = \text{Attn}([\text{RRCP}(c_i, c_{i,k}^r) \cdot \mathbf{z}_i^v]_{i \in [1, |C_i^s|]}), \mathbf{z}_i^v \in \mathbf{Z}_i^v, \quad (6)$$

$$\mathbf{z}'_{i,t} = \text{Attn}([\text{RRCP}(c_i, c_{i,k}^r) \cdot \mathbf{z}_i^t]_{i \in [1, |C_i^s|]}), \mathbf{z}_i^t \in \mathbf{Z}_i^t. \quad (7)$$

Dataset	ICIP	SMPD	Instagram
# UGCs	20,337	305,613	297,865
# Users	17,302	38,312	33,935
avg. UGC Popularity	200.78	493.14	4,694.26
avg. Text Length	27.68	91.75	442.78

Table 1: Data Statistics

At last, FC layers are used to predict the UGC popularity and mean squared error is used as the optimization loss:

$$\hat{p}_i = \text{FC}(\mathbf{z}'_{i,v} \oplus \mathbf{z}'_{i,t}), \quad \mathcal{L}_{\text{loss}} = \frac{1}{N} \sum_{i=1}^{N} (p_i - \hat{p}_i)^2. \quad (8)$$

4 Experiments

We evaluate the proposed SKAPP model for multimodal SMPP task against state-of-the-arts methods. Our experiments span three social media UGC datasets, compare eleven strong baselines, and include ablation studies, parameter and complexity analyses, and robustness.

Datasets Three real-world social media datasets comprising multimodal UGCs: **ICIP** (Ortis, Farinella, and Battiato 2019), **SMPD** (Wu et al. 2023), and **Instagram** (Kim et al. 2020). Table 1 presents the basic statistics of datasets.

Baselines We compare SKAPP with the following eleven baselines. They include three feature engineering-based approaches: SVR (Khosla, Das Sarma, and Hamid 2014), HyFea (Lai, Zhang, and Zhang 2020), and MFTM (Hsu et al. 2023); Six deep learning approaches: CLSTM (Ghosh et al. 2016), HMMVED (Xie, Zhu, and Chen 2023), DLBA (Brunelli, Viola, and Susto 2021), MASSL (Zhang et al. 2022), BLIP (Li et al. 2022), and CBAN (Cheung and Lam 2022); Two retrieval-based approaches: NIPA (Ji et al. 2023a) and NMRA (Zhong et al. 2024).

Metrics Following existing works (Cappallo, Mensink, and Snoek 2015; Wu et al. 2023), three standard metrics were used: mean squared error (MSE), mean absolute error (MAE), and Spearman's rank correlation (SRC).

Implementation The dataset split ratio is 8:1:1 for training, validation, and test sets, respectively. We use PyTorch to implement the SKAPP model, with Adam optimizer and an initial learning rate of $1e^{-4}$. The number of retrieved UGCs is 500, k_1 of BM25 is 0.5, the threshold θ of selective refiner is 0, and the dimensions for embeddings $\mathbf{v}$ and $\mathbf{t}$ are 768.

Main Results

Table 2 presents the prediction performance of our model compared to eleven baselines on the SMPP task. The following observations can be made: (i) Feature-engineering models generally perform on par with deep learning models, although their performance can sharply decline in certain cases, as seen with the SVR model on the SMPD dataset; (ii) Among the six deep learning models, the performance of CLSTM, HMMVED, and CBAN is higher than that of the DLBA, MASSL, and BLIP; (iii) Of the two

Method	Type	ICIP			SMPD			Instagram		
		MSE	MAE	SRC	MSE	MAE	SRC	MSE	MAE	SRC
SVR	Feature	1.9009	0.8941	0.5241	6.2996	2.0208	0.2163	7.0534	1.9695	0.4035
HyFea	Feature	1.9013	1.0181	0.4497	4.7429	1.7080	0.4677	4.7132	1.6924	0.4708
MFTM	Feature	1.8970	0.9772	0.4156	4.0222	1.5481	0.5849	4.3073	1.6132	0.5321
CLSTM	Deep	1.8724	0.9823	0.4654	3.9143	1.5005	0.5888	4.2431	1.5882	0.5396
HMMVED	Deep	1.8556	0.9497	0.4524	3.7154	1.3636	0.6352	4.2461	1.6017	0.5385
DLBA	Deep	2.2290	1.0097	0.3614	4.8693	1.7021	0.4387	5.1425	1.7527	0.4007
MASSL	Deep	1.9446	0.9278	0.4499	5.5670	1.8427	0.5271	7.8583	2.2274	0.5188
BLIP	Deep	2.0646	0.9961	0.3603	4.3884	1.6340	0.5269	5.2436	1.8058	0.3762
CBAN	Deep	1.8098	0.9309	0.4727	4.0443	1.5123	0.5754	4.2808	1.5894	0.5426
NIPA	Retrieval	1.9999	0.9980	0.3989	4.2538	1.6532	0.4086	4.0209	1.5565	0.5696
MMRA	Retrieval	1.7600	0.8684	0.5439	3.5119	1.3730	0.6423	3.9456	1.5070	0.5806
SKAPP	Retrieval	**0.9662**	**0.6367**	**0.6965**	**1.8196**	**0.8249**	**0.8414**	**2.0936**	**1.0369**	**0.8272**
(improv.)		39.61%↑	26.68%↑	28.06%↑	48.19%↑	39.51%↑	31.00%↑	46.94%↑	29.06%↑	42.47%↑

Table 2: Social media popularity prediction performance comparison between our proposed SKAPP model and eleven baselines on three large-scale real-world datasets. The best results are marked in bold and the second best are underlined.

retrieval-based models, NIPA underperforms on the ICIP and SMPD datasets. This may be because it mainly considers the visual semantic similarity in retrieval, neglecting textual and metadata information; In contrast, MMRA, which employs both UGC photos and texts in retrieval, achieves the best performance compared to baselines; (iv) Our proposed SKAPP model, powered with a meta retriever for retrieving more diverse and relevant UGCs, a selective refiner for distilling retrieved UGCs based on the RRCP measure, and a VL-GNN prediction network for aggregating contextual UGC knowledge, remarkably outperforms all baselines. It achieves relative improvements in MSE of up to 39.61%, 48.19%, and 46.94% on the ICIP, SMPD, and Instagram datasets, respectively, compared to the second best results.

Experimental Analysis

The results of ablation studies for SKAPP's modules, UGC modalities, and retrieving strategies are presented in Table 3.

Ablation Study To investigate the contributions of M's key modules, we create five variant models by removing one module at a time. The ablation results reveal the following: (i) without the retrieval module, our model's performance is in line with the baselines, highlighting the critical role of retrieval in improving SMPP performance; (ii) when only semantic retrieval is used (i.e., without the meta retriever), the model's performance significantly declines. This is likely because semantic retrieval can introduce a large number of similar but irrelevant UGCs; (iii) the selective refiner is essential for effective retrieval, as it selects UGCs that positively contribute to the prediction; (iv) the VL-GNN module enhances performance by aggregating multimodal and contextual UGC knowledge through graph learning; (v) the RRCP-Attention prediction network further boosts performance by applying two levels of weighted fusion to the learned UGC representations; (vi) overall, the integration of all modules yields the lowest prediction error across the three datasets, demonstrating the effectiveness of our proposed modules in SKAPP.

Variant	ICIP	SMPD	Instagram
Ablation of SKAPP's Modules			
w/o Retrieval	1.5614	4.0443	3.2734
w/o Meta Retriever	1.9006	4.1353	5.2537
w/o Selective Refiner	1.1004	2.0854	2.6332
w/o VL-GNN	1.1223	2.1056	2.7178
w/o RRCP-Attention	1.0761	1.9606	2.1636
Ablation of UGC modalities			
w/o Visual	1.1770	2.3567	2.4851
w/o Textual	1.1829	2.7037	2.3582
w/o Metadata	1.8188	4.0359	5.2537
Ablation of Retrieving Strategies			
retrieval based on Photo	1.9006	4.1353	5.7644
retrieval based on Texts	1.9653	3.9958	4.7259
retrieval based on Metadata	1.6280	2.6945	3.8679
retrieval based on FLICO	1.8255	3.8562	5.4786
retrieval based on NIPA	1.9321	4.1687	5.2468
retrieval based on MMRA	1.9627	4.0507	4.6693
SKAPP (Full)	0.9662	1.8196	2.0936

Table 3: Ablation studies on SKAPP's modules, UGC modalities, and retrieving strategies across three datasets. The evaluation metric is MSE.

UGC Modality We further ablate the effects of three UGC modalities by excluding one modality – visual, textual, or metadata – from SKAPP's input. The results indicate that metadata is the most important modality, supporting our design of incorporating metadata into the retrieval query construction. Both visual and textual modalities contribute non-trivially to the prediction, with the visual modality being more influential on the SMPD dataset, and the textual modality playing a more vital role on the Instagram dataset.

Retrieval Strategy To evaluate the effectiveness of the proposed meta retriever and selective refiner, we compare SKAPP with several retrieval strategies: retrieval based on photos, texts, metadata, and three strategies employed in

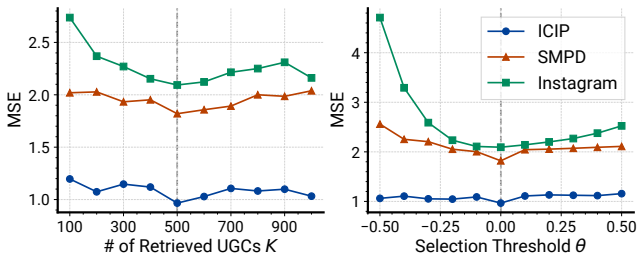

Figure 4: Parameter Analysis of SKAPP

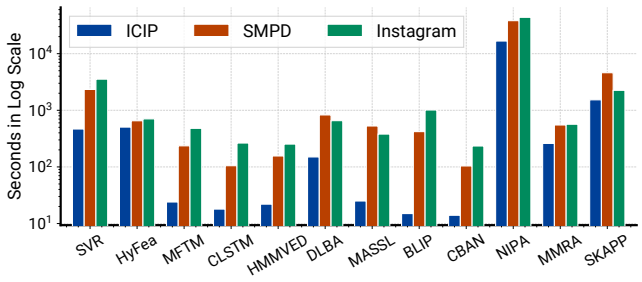

Figure 5: Training Time of Baselines and SKAPP

FLICO (Wang et al. 2023), NIPA (Ji et al. 2023a), and MMRA (Zhong et al. 2024). We can see that the semantic retrieval strategies – whether based on photo, texts, NIPA (photo), or MMRA (photo and texts) – are less effective than metadata-based retrieval strategies. FLICO's approach, which uses a separate classification model to estimate the predictive contributions of UGCs, is ineffective for complex social UGCs. These findings validate the efficacy of our proposed meta retriever and selective refiner.

Parameter Sensitivity We perform a sensitivity analysis on two key hyperparameters of SKAPP: the number of retrieved UGCs K and the contribution threshold θ of RRCP. As shown in Figure 4, the optimal performance is generally achieved with $K \approx 500$ and $\theta \approx 0$.

Complexity Analysis The computational complexity of SKAPP primarily arises from the retrieval process and the prediction network. The embedding extraction and image-to-text transformation of the UGCs in $\mathcal{D}$ are performed in advance, incurring no real-time costs during UGC prediction. The complexity of BM25 is $\mathcal{O}(|Q|*|\mathcal{D}|+|\mathcal{D}|\times\log|\mathcal{D}|)$. The complexity of the VL-GNN is mainly determined by the number of nodes $|\mathcal{V}| = K = 500$ in $\mathcal{G}$, which is fixed. We compared SKAPP's training time with that of the baselines, as shown in Figure 5. NIPA is the most computational intensive model, followed by SKAPP. Although SKAPP is not as lightweight as models like CLSTM or CBAN, its substantial performance gains justify the additional computational overhead. In practice, for a dataset of approximately 300K UGCs, the prediction and retrieval costs of SKAPP are about 50 seconds and 7 hours, respectively, when running on a system with a 5.40GHz CPU, an NVIDIA 3090Ti GPU with 24GB memory, and 24GB DDR4 RAM at 3200MHz. A comparison of the retrieval and prediction times for the three retrieval-based models is shown in Table 4. MMRA's retrieval process is more efficient due to its smaller number of query features compared to NIPA and SKAPP. It is also worth noting that the retrieval process can be further accelerated through multi-threaded processing.

Model	ICIP		SMPD		Instagram	
	Retr.	Pred.	Retr.	Pred.	Retr.	Pred.
NIPA	4.9m	53.6s	9.0h	18.3m	10.5h	14.5m
MMRA	0.9m	4.5s	1.8h	3.9s	2.1h	3.7s
SKAPP	9.5m	41.2s	7.0h	50.1s	7.3h	45.0s

Table 4: Time Comparison for Retrieval and Training

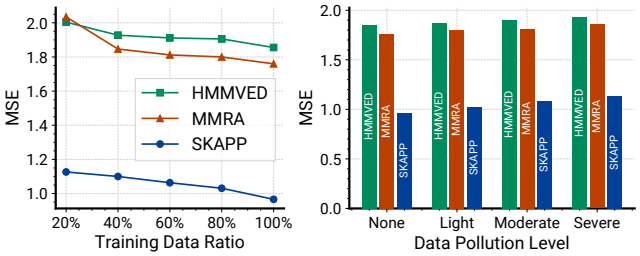

Figure 6: Robustness Under Data Sparsity and Pollution

Robustness To evaluate the robustness of SKAPP, we conduct experiments under two conditions: (i) with reduced training data; and (ii) with polluted UGCs and retrieval results. For UGC pollution, we introduce Gaussian noise to the visual and textual UGC embeddings. For retrieval pollution, we randomly replace retrieved UGCs with irrelevant ones from the database. We define three levels of pollution – light, moderate, and severe, corresponding to 0.1, 0.3, and 0.5, respectively, in the variances of the Gaussian noise and in the replacement ratios. We compare SKAPP with two top-performing baselines, HMMVED and MMRA, on the ICIP dataset. The results, presented in Figure 6, show that all three models experience performance degradation with reduced trained data or polluted data. The impact of training data size on performance is more pronounced than that of data quality. Despite these challenges, SKAPP consistently outperforms the baselines by large margins, demonstrating its robustness in scenarios of data sparsity and data pollution.

5 Conclusion

In this work, we studied the multimodal SMPP task and proposed SKAPP, a retrieval-based knowledge augmentation framework. Our approach includes the design of a meta retriever that queries a diverse set of potentially relevant UGCs and a selective refiner that retains UGCs with positive gains in prediction. Moreover, we employed a VL-GNN and an RRCP-Attention-based prediction network to aggregate the retrieved knowledge. Experiments on three datasets demonstrated the effectiveness and robustness of SKAPP in enhancing SMPP performance. Future work can explore ways to improve the retrieval efficiency, construct better queries, and design new selection strategies.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant Nos. U22A2097, 62072077, and 62186043.

References

Abbar, S.; Castillo, C.; and Sanfilippo, A. 2018. To post or not to post: Using online trends to predict popularity of offline content. In *ACM HT*, 215–219.

Aggrawal, N.; Ahluwalia, A.; Khurana, P.; and Arora, A. 2017. Brand analysis framework for online marketing: Ranking web pages and analyzing popularity of brands on social media. *Social Network Analysis and Mining*, 7: 1–10.

Brunelli, L.; Viola, M.; and Susto, G. A. 2021. Instagram Images and Videos Popularity Prediction: A Deep Learning-Based Approach. In *Italian Workshop on Artificial Intelligence and Applications for Business and Industries*.

Cao, Q.; Shen, H.; Gao, J.; Wei, B.; and Cheng, X. 2020. Popularity prediction on social platforms with coupled graph neural networks. In *WSDM*, 70–78.

Cappallo, S.; Mensink, T.; and Snoek, C. G. 2015. Latent factors of visual popularity prediction. In *ICMR*, 195–202.

Chen, X.; Chen, W.; Huang, C.; Zhang, Z.; Duan, L.; and Zhang, Y. 2023. Double-Fine-Tuning Multi-Objective Vision-and-Language Transformer for Social Media Popularity Prediction. In *ACM MM*, 9462–9466.

Cheng, Z.; Walker, J.; Zhong, T.; and Zhou, F. 2022. Modeling multi-view interactions with contrastive graph learning for collaborative filtering. In *IJCNN*.

Cheng, Z.; Zhou, F.; Xu, X.; Zhang, K.; Trajcevski, G.; Zhong, T.; and Philip, S. Y. 2024. Information Cascade Popularity Prediction via Probabilistic Diffusion. *TKDE*, 36(12): 8541–8555.

Cheung, T.-h.; and Lam, K.-m. 2022. Crossmodal bipolar attention for multimodal classification on social media. *Neurocomputing*, 514: 1–12.

Cuconasu, F.; Trappolini, G.; Siciliano, F.; Filice, S.; Campagnano, C.; Maarek, Y.; Tonellotto, N.; and Silvestri, F. 2024. The power of noise: Redefining retrieval for RAG systems. In *SIGIR*, 719–729.

Fernandes, P.; Yin, K.; Neubig, G.; and Martins, A. F. 2021. Measuring and Increasing Context Usage in Context-Aware Machine Translation. In *ACL*, 6467–6478.

Ferrara, E.; Interdonato, R.; and Tagarelli, A. 2014. Online popularity and topical interests through the lens of Instagram. In *ACM HT*, 24–34.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997.

Ghosh, S.; Vinyals, O.; Strope, B.; Roy, S.; Dean, T.; and Heck, L. 2016. Contextual LSTM (CLSTM) models for large scale NLP tasks. arXiv:1602.06291.

Gu, J.; Xu, X.; Tian, Y.; Hu, Y.; Huang, J.; Zhong, W.; Zhou, F.; and Gao, L. 2024. RRE: A Relevance Relation Extraction Framework for Cross-domain Recommender System at Alipay. In *ICME*, 1–6.

Hong, L.; Dan, O.; and Davison, B. D. 2011. Predicting popular messages in Twitter. In *WWW*, 57–58.

Hsu, C.-C.; Lee, C.-M.; Hou, X.-Y.; and Tsai, C.-H. 2023. Gradient Boost Tree Network based on Extensive Feature Analysis for Popularity Prediction of Social Posts. In *ACM MM*, 9451–9455.

Ji, L.; Park, C. H.; Rao, Z.; and Chen, Q. 2023a. Neural Image Popularity Assessment with Retrieval-augmented Transformer. In *ACM MM*, 2427–2436.

Ji, S.; Lu, X.; Liu, M.; Sun, L.; Liu, C.; Du, B.; and Xiong, H. 2023b. Community-based dynamic graph learning for popularity prediction. In *ACM KDD*, 930–940.

Kevin Wu, J. Z., Eric Wu. 2024. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. arXiv:2404.10198v2.

Khosla, A.; Das Sarma, A.; and Hamid, R. 2014. What makes an image popular? In *WWW*, 867–876.

Kim, S.; Jiang, J.-Y.; Nakada, M.; Han, J.; and Wang, W. 2020. Multimodal Post Attentive Profiling for Influencer Marketing. In *WWW*, 2878–2884.

Lai, X.; Zhang, Y.; and Zhang, W. 2020. HyFea: Winning solution to social media popularity prediction for multimedia grand challenge 2020. In *ACM MM*, 4565–4569.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 12888–12900.

Li, X.; Lian, D.; Lu, Z.; Bai, J.; Chen, Z.; and Wang, X. 2024. GraphAdapter: Tuning vision-language models with dual knowledge graph. In *NeurIPS*, 13448–13466.

Lin, W.; Chen, J.; Mei, J.; Coca, A.; and Byrne, B. 2024. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. In *NeurIPS*, 22820–22840.

Liu, A.-A.; Wang, X.; Xu, N.; Guo, J.; Jin, G.; Zhang, Q.; Tang, Y.; and Zhang, S. 2022. A review of feature fusion-based media popularity prediction methods. *Visual Informatics*, 6(4): 78–89.

Long, X.; Zeng, J.; Meng, F.; Ma, Z.; Zhang, K.; Zhou, B.; and Zhou, J. 2024. Generative multi-modal knowledge retrieval with large language models. In *AAAI*, 18733–18741.

Mishra, S.; Rizoiu, M.-A.; and Xie, L. 2016. Feature driven and point process approaches for popularity prediction. In *CIKM*, 1069–1078.

Momeni, E.; Cardie, C.; and Diakopoulos, N. 2015. A survey on assessment and ranking methodologies for user-generated content on the web. *ACM Computing Surveys*, 48(3): 1–49.

Moniz, N.; and Torgo, L. 2019. A review on web content popularity prediction: Issues and open challenges. *Online Social Networks and Media*, 12: 1–20.

Ortis, A.; Farinella, G. M.; and Battiato, S. 2019. Prediction of social image popularity dynamics. In *ICIAP*, 572–582.

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M. M.; Gatford, M.; et al. 1995. Okapi at TREC-3. *NIST Special Publication*, 109: 109.

Tatar, A.; De Amorim, M. D.; Fdida, S.; and Antoniadis, P. 2014. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5: 1–20.

Wang, Z.; Araki, J.; Jiang, Z.; Parvez, M. R.; and Neubig, G. 2023. Learning to filter context for retrieval-augmented generation. arXiv:2311.08377.

Wu, B.; Liu, P.; Cheng, W.-H.; Liu, B.; Zeng, Z.; Wang, J.; Huang, Q.; and Luo, J. 2023. SMP Challenge: An Overview and Analysis of Social Media Prediction Challenge. In *ACM MM*, 9651–9655.

Xie, J.; Zhu, Y.; and Chen, Z. 2023. Micro-video popularity prediction via multimodal variational information bottleneck. *IEEE Transactions on Multimedia*, 25: 24–37.

Xu, X.; Zhou, F.; Zhang, K.; Liu, S.; and Trajcevski, G. 2023. CasFlow: Exploring hierarchical structures and propagation uncertainty for cascade prediction. *TKDE*, 35(4): 3484–3499.

Zhang, Y.; Feng, F.; He, X.; Wei, T.; Song, C.; Ling, G.; and Zhang, Y. 2021. Causal intervention for leveraging popularity bias in recommendation. In *SIGIR*, 11–20.

Zhang, Z.; Xu, S.; Guo, L.; and Lian, W. 2022. Multi-modal Variational Auto-Encoder Model for Micro-video Popularity Prediction. In *ICCIP*, 9–16.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *ACM KDD*, 1513–1522.

Zhong, T.; Lang, J.; Zhang, Y.; Cheng, Z.; Zhang, K.; and Zhou, F. 2024. Predicting Micro-video Popularity via Multimodal Retrieval Augmentation. In *SIGIR*, 2579–2583.

Zhou, F.; Xu, X.; Trajcevski, G.; and Zhang, K. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys*, 54(2): 1–36.