Learning to Curate Context: Jointly Optimizing Retrieval and Prediction for Multimodal Social Media Popularity

Xovee Xu, Shuojun Lin, Fan Zhou, Jingkuan Song*

University of Electronic Science and Technology of China
Chengdu, Sichuan 610054, China
xovee.xu@gmail.com, locklin0223@gmail.com, fan.zhou@uestc.edu.cn, jingkuan.song@gmail.com

Abstract

Predicting the popularity of user-generated content (UGC) is a crucial but challenging task in social media analysis. While existing retrieval-augmented models enhance predictions by supplying rich contextual information, they remain limited by a fundamental precision-recall dilemma: enlarging the retrieval set increases coverage but introduces noisy, irrelevant context that harms prediction. In this work, we propose a unified framework that learns to retrieve, filter, and predict. Central to our approach is a Mixture-of-Logits-based retrieval module that replaces static similarity metrics with a dynamic, multi-faceted scoring function, enabling the retriever to be directly optimized by the prediction objective. Then an uncertainty-aware filter is designed to perform differentiable subset selection and refine the selected representations using the information bottleneck principle. At last, to enhance predictive robustness, we introduce a confidence-weighted test-time perturbation strategy. By learning to retrieve UGCs that are beneficial for prediction and filtering out uncertainty, our framework provides more relevant and reliable context. Extensive experiments demonstrate that the proposed framework achieves state-of-the-art performance, consistently outperforming strong baselines.

1 Introduction

Predicting the popularity of user-generated content (UGC) represents a cornerstone task in modern social network analysis with broad scientific and practical value (Tatar et al. 2014). The core challenge stems from UGC's informality, noise, and heterogeneity, where engagement outcomes are often stochastic and shaped by network effects, exogenous events, and platform algorithms (Naab and Sehl 2017). Consequently, effective prediction demands models that can jointly reason about the intrinsic quality of multimodal content—spanning text, images, and video—and the extrinsic social context that governs information dissemination among users (Hsu et al. 2024). Reliable solutions enable key applications: improved recommendation and ad allocation for platforms (Jeon et al. 2024; Gu et al. 2024), decision support and personalization for users and creators (Zhou et al. 2024), and societal benefits such as timely public-interest messaging and early detection of harmful or misleading virality (Sun et al. 2025; Drolsbach and Pröllochs 2023; Lang et al. 2025).

Prior work on UGC popularity prediction in social networks largely follows two streams: content-centric and network/temporal. *Content-centric* models focus on the intrinsic appeal of UGC itself, evolving from feature-engineering models (Cheng et al. 2014) to deep architectures (CNNs, Transformers) that learn joint multimodal representations (Xie et al. 2020). By contrast, *network/temporal* models prioritize extrinsic diffusion dynamics, employing temporal point processes (Zhao et al. 2015), survival analysis (Gao et al. 2020), and graph neural networks (GNNs) over follower and content interaction graphs to capture contagion, seasonality, exogenous shocks, and user influence (Li et al. 2021; Zhou et al. 2021; Xu et al. 2021, 2022). Content models scale when only the UGC is available but degrade when exposure mechanisms dominate or trends shift; network/temporal methods excel once early signals or social context are present yet depend on high-quality graphs (often unavailable) and may generalize poorly across platforms or time. Across both streams, accurately learning intrinsic content quality and user preferences—and reliably estimating diffusion paths, temporal momentum, and network topology (Cao et al. 2020; Ji et al. 2023b)—are critical yet challenging due to multimodal noise, sparse or partial observations, and rapidly evolving contexts.

Recently, retrieval-based approaches have emerged as a promising alternative (Cheng et al. 2024). They retrieve relevant UGCs from a large corpus as knowledge augmentations and condition the predictor on these "neighbors." By grounding popularity prediction in large-scale semantic and social contexts, retrieval-based methods have achieved significant performance gains without relying on explicit network structures. For example, NIPA (Ji et al. 2023a) and MMRA (Zhong et al. 2024) models retrieve relevant UGCs as contextual information to enhance the learning of the target UGC; The SKAPP (Xu et al. 2025) introduces a meta retriever to find UGCs that are similar not only semantically but also in terms of user profiles and posting dynamics.

However, the effectiveness of existing retrieval-based approaches is hindered by two key limitations. (1) *Misaligned retrieval and prediction*: The retrieval process often relies on proxy objectives (e.g., embedding similarity and scalar scores) rather than being guided by the end task targets.

*Corresponding author.

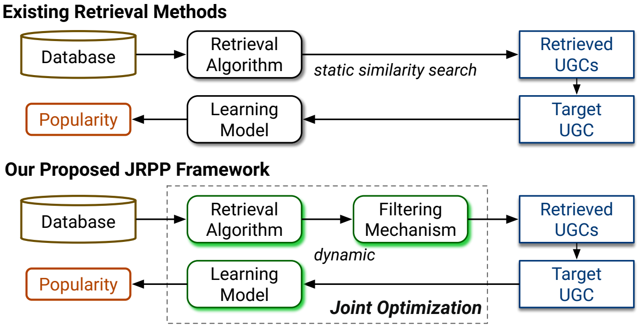

Figure 1: From static retrieval to jointly optimized, task-aligned social media UGC popularity prediction.

This misalignment surfaces off-topic or redundant retrievals, thereby increasing noise in the augmented context and diminishing the predictive gains from retrieval. (2) *Sensitivity to retrieval noise*: Current models are highly susceptible to retrieval noise and the stochastic nature of social media engagement, as they rarely quantify the predictive quality and ambiguity of retrieved UGCs. Although a few works have proposed post-hoc selection mechanisms (Xu et al. 2025), these are not jointly optimized with UGC representation learning and aggregation, inevitably introducing high variance and poor calibration under distribution shifts.

These limitations motivate a new retrieval-based framework that learns task-aware retrieval and uncertainty-aware context integration end-to-end, ensuring that the retrieved knowledge aligns with the prediction objective. Specifically, we propose JRPP, Jointly optimized Retrieval and Prediction for multimodal social media Popularity. Diverging from the traditional retrieval-based approaches (see a comparison in Figure 1), JRPP integrates retrieval and prediction into a unified, end-to-end trainable architecture. Our solution consists of three core modules. First, we introduce a Mixture-of-Logits (MoL) based retrieval mechanism that replaces static similarity metrics with a dynamic, multi-faceted scoring function, enabling the retriever to be directly optimized by the final prediction objective. Second, to mitigate the impact of retrieval noise, we design a two-stage uncertainty-aware filter, which first selects a high-quality subset of retrieved UGCs via a differentiable Gumbel-Softmax gate and then refines their representations using the information bottleneck principle to isolate task-relevant information. At last, we employ a confidence-weighted test-time perturbation strategy to enhance the model's robustness against UGC variations.

Our main contributions are summarized as follows:

- We introduce the first *joint retrieval and prediction* framework for UGC popularity prediction. JRPP optimizes the retrieval module and the learning model in a single loop, eliminating objective mismatch and enabling retrieval to co-adapt with the prediction objective.
- We design a MoL-based retrieval module for expressive, task-aligned matching; an uncertainty-aware filter that couples differentiable subset selection with an IB-based refinement; and a test-time perturbation strategy that self-ensembles predictions for improved robustness.
- Extensive experiments on three large-scale social UGC datasets show that JRPP achieves state-of-the-art performance, outperforming strong baselines, up to 25.36% on MSE, 18.29% on MAE, and 5.22% on SRC. Additional experiments including ablation study and parameter analysis further validate the effectiveness of our model.

2 Preliminaries

Problem Definition Given a dataset of N multimodal UGCs $C = \{c_i\}_{i=1}^{N}$, each UGC $c_i = \{t_i, v_i, u_i, \dots\}$ is composed of multiple data modalities such as text t_i, image v_i, and user u_i. The goal of multimodal UGC popularity prediction is to learn a model $\mathcal{M}$ that takes a UGC c_i as input and predict its future popularity p_i:

$$\hat{p}_i = \mathcal{M}(c_i), \tag{1}$$

where $\hat{p}_i$ is the predicted future popularity of c_i. The definition of popularity varies across contexts, e.g., the numbers of likes and reshares of social tweets.

Retrieval-based Popularity Prediction To enhance predictive performance, retrieval-based approaches leverage contextual information from a large database $\mathcal{D}$. Instead of relying solely on the target UGC c_i, the model's input is enriched with a set of k relevant items, $\mathcal{R}_i = \{r_1, r_2, \dots, r_k\}$, retrieved from $\mathcal{D}$. These items (UGCs, in our case) are selected using a scoring function $s(c_i, r)$ that measures their relevance (e.g., semantic or social similarity) to the target c_i. An aggregation function $A(\cdot,\cdot)$ then fuses the information from c_i and the retrieved set $\mathcal{R}_i$. Here $\mathcal{R}_i$ is defined as the top-k UGCs according to $s(c_i, r)$, equivalently:

$$\mathcal{R}_i = \underset{S \subseteq \mathcal{D}:\, |S|=k}{\arg\max} \sum_{r \in S} s(c_i, r). \tag{2}$$

The final popularity prediction is defined as:

$$\hat{p}_i = \mathcal{M}\left(A(c_i, \mathcal{R}_i)\right), \tag{3}$$

where $\hat{p}_i$ is the predicted popularity of the target UGC c_i.

3 Methodology

Overview Our proposed JRPP framework introduces an end-to-end framework that jointly optimizes UGC retrieval and popularity prediction. It features three key modules: a Mixture-of-Logits-based retrieval module for expressive, task-aligned matching; an uncertainty-aware filter that selects and refines retrieved content to reduce noise; and a test-time perturbation method to enhance prediction robustness. Figure 2 presents a framework overview.

Joint Retrieval and Prediction by Mixture of Logits

Previous retrieval-based models design the retrieval module with surrogate objectives (e.g., semantic and social similarity) (Zhong et al. 2024; Xu et al. 2025) and then freeze it while a separate learning model aggregates the retrieved UGCs with the query UGC. This separation causes a persistent objective mismatch: the retrieved UGCs are not aligned

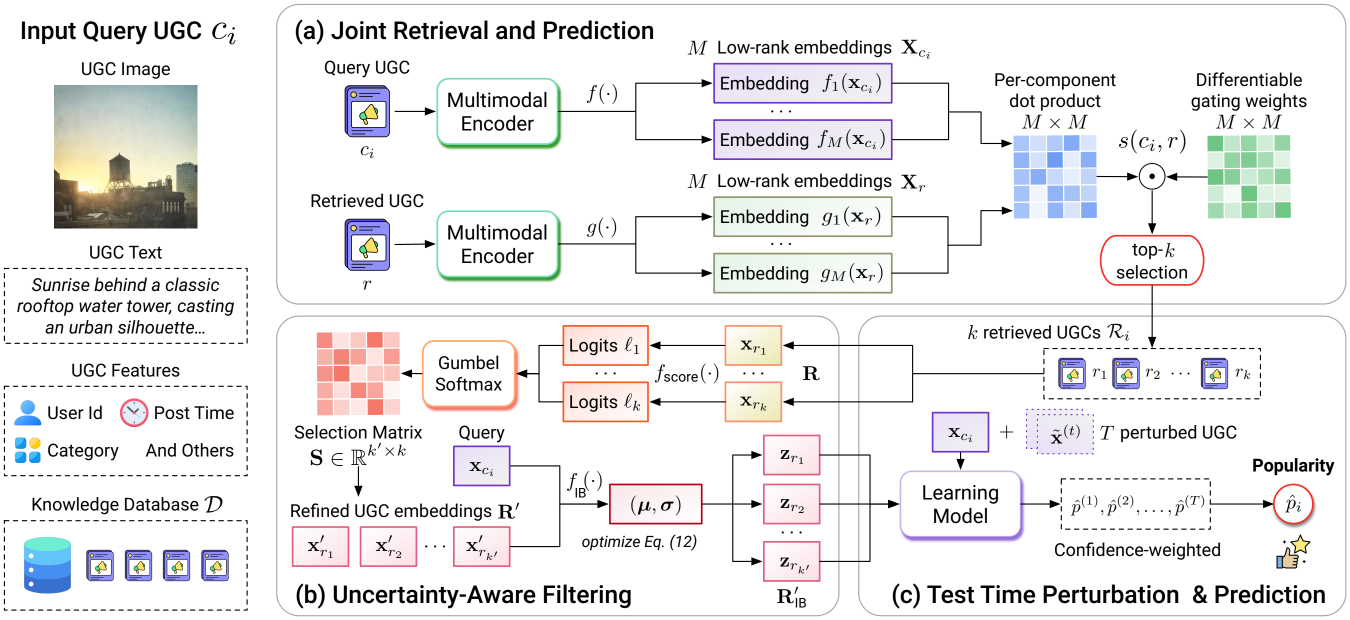
Input Query UGC c_i
UGC Image
UGC Text
Sunrise behind a classic rooftop water tower, casting an urban silhouette...
UGC Features
User Id
Post Time
Category
And Others
Knowledge Database $\mathcal{D}$
(a) Joint Retrieval and Prediction
Query UGC
c_i
Multimodal Encoder
$f(\cdot)$
M Low-rank embeddings $\mathbf{X}_{c_i}$
Embedding $f_1(\mathbf{x}_{c_i})$
Embedding $f_M(\mathbf{x}_{c_i})$
Retrieved UGC
r
Multimodal Encoder
$g(\cdot)$
M Low-rank embeddings $\mathbf{X}_r$
Embedding $g_1(\mathbf{x}_r)$
Embedding $g_M(\mathbf{x}_r)$
Per-component dot product $M \times M$
$s(c_i, r)$
Differentiable gating weights $M \times M$
top-k selection
(b) Uncertainty-Aware Filtering
Selection Matrix $\mathbf{S} \in \mathbb{R}^{k' \times k}$
Gumbel Softmax
Logits ℓ_1
Logits ℓ_k
$f_{\text{score}}(\cdot)$
$\mathbf{x}_{r_1}$
$\mathbf{x}_{r_k}$
$\mathbf{R}$
Query
$\mathbf{x}_{c_i}$
$f_{\text{IB}}(\cdot)$
$(\boldsymbol{\mu}, \boldsymbol{\sigma})$
optimize Eq. (12)
Refined UGC embeddings $\mathbf{R}'$
$\mathbf{x}'_{r_1}$
$\mathbf{x}'_{r_2}$
$\mathbf{x}'_{r_{k'}}$
$\mathbf{z}_{r_1}$
$\mathbf{z}_{r_2}$
$\mathbf{z}_{r_{k'}}$
$\mathbf{R}'_{\text{IB}}$
(c) Test Time Perturbation & Prediction
k retrieved UGCs $\mathcal{R}_i$
r_1
r_2
r_k
$\mathbf{x}_{c_i}$
+
$\tilde{\mathbf{x}}^{(t)}$
T perturbed UGC
Learning Model
$\hat{p}^{(1)}, \hat{p}^{(2)}, \ldots, \hat{p}^{(T)}$
Confidence-weighted
Popularity
$\hat{p}_i$

Figure 2: Overview of the proposed JRPP framework. It takes the query UGC as input, retrieves relevant UGCs from a database, and predicts query UGC's future popularity. It consists of three key modules: (a) joint retrieval and prediction; (b) uncertainty-aware retrieval filtering; and (c) test-time perturbations. For clarity, only one retrieved UGC is illustrated.

with what ultimately matters—the accurate popularity prediction, in our case—and improvements to the learning model cannot feed back to shape the retriever (via gradient backpropagation for deep neural networks).

To enable jointly optimized retrieval and prediction, we need not only a scoring function that is parameterized and trained end-to-end under the popularity prediction loss, but also a more expressive retrieval mechanism that is adapted for our joint optimization framework. Inspired by the Mixture-of-Logits (MoL) (Zhai et al. 2023; Ding and Zhai 2025), we propose a new retrieval framework to jointly perform retrieval and popularity prediction for social media UGCs. Without loss of generality, given an arbitrary pair of a query UGC c_i and a candidate UGC r, we first project the two UGCs' multimodal content (text, image, and features from UGC metadata and users) into two embeddings $\mathbf{x}_{c_i}, \mathbf{x}_r \in \mathbb{R}^d$ via pre-trained multimodal encoders.

Instead of directly assessing the similarity between $\mathbf{x}_{c_i}$ and $\mathbf{x}_r$ via a single dot-product, we project the input embeddings to a mixture of M pairs of low-rank embeddings:

$$\mathbf{X}_{c_i} = \{f_m(\mathbf{x}_{c_i})\}_{m=1}^{M}, \quad \mathbf{X}_r = \{g_m(\mathbf{x}_r)\}_{m=1}^{M}, \tag{4}$$

where each low-rank embedding $f_m(\mathbf{x})$ and $g_m(\mathbf{x})$ is processed by a projection layer ($f_m(\cdot)$ for query and $g_m(\cdot)$ for candidate) with nonlinearity and an ℓ_2 normalization. The scoring function $s(c_i, r)$ is then defined by a gated sum of per-component cosine scores:

$$s(c_i, r) = \sum_{m=1}^{M} w_m(c_i, r) \frac{f_m(c_i)^\top g_m(r)}{\|f_m(c_i)\| \|g_m(r)\|}, \tag{5}$$

where $w_m(c_i, r) \in [0, 1]$ is the adaptive gating weights, initialized by fully-connected (FC) layers. This design allows the learning model to dynamically emphasize or suppress individual low-rank embeddings per query-candidate UGC pair while retaining the theoretical capacity to overcome the single-vector low-rank bottleneck and approximate any full-rank similarity matrix. This closed-loop optimization yields task-aligned retrieval, better calibration of scores, and more robust behavior under distribution shift, enabling conditional, multi-component matching that co-adapts with the downstream learning model.

Let $\mathcal{R}_i = \{r_1, r_2, \ldots, r_k\}, k \ll |\mathcal{D}|$ denote the set of k retrieved UGCs for a query UGC c_i from the database $\mathcal{D}$, ranked by the calculated scores $s(c_i, r)$. We can aggregate the embeddings of the target and retrieved UGCs by a Transformer-based learning model $\mathcal{M}$ for the popularity prediction task:

$$\hat{p}_i = \mathcal{M}\left([\mathbf{x}_{c_i}; \mathbf{x}_{r_1}; \mathbf{x}_{r_2}; \ldots; \mathbf{x}_{r_k}]\right). \tag{6}$$

Apart from the discrete top-k operator used during retrieval, the learning model and the gated weights are differentiable and trained jointly. Next, we introduce an uncertainty-aware filtering mechanism that further improves the quality of $\mathcal{R}_i$.

Uncertainty-Aware Retrieval Filtering

Social UGC is notoriously noisy, heterogeneous, and contextually diverse: posts differ wildly in style, length, language and image quality, and topical focus (Krumm, Davies, and Narayanaswami 2008). For a retrieval-based social media popularity prediction model, the quality and relevance of the retrieved UGCs significantly impact the prediction results. The complexity of social UGCs could make a naïve top-k retrieval vulnerable to returning irrelevant or misleading UGCs that degrade downstream popularity prediction. Even when the retrieval framework is trained jointly with the predictor, the presence of low-quality candidates in $\mathcal{R}_i$ can inject noise into the aggregated representation and force the model to allocate capacity to denoise the retrievals—capacity that could instead be used to learn fine-grained popularity cues—and amplifies distribution shift risks at inference time. To address this, we design a two-stage, uncertainty-aware filtering pipeline. The first stage performs a differentiable subset selection using Gumbel-Softmax mechanism, and the second stage then refines the representations of the selected UGCs using the theory of information bottleneck (IB) (Tishby, Pereira, and Bialek 2000; Tishby and Zaslavsky 2015; Zhu et al. 2024). This cascaded approach first selects a high-quality subset of candidates and then compresses their representations to preserve maximal information with the popularity labels, while discarding task-irrelevant variability.

Differentiable UGC Selection via Gumbel-Softmax The Gumbel-Softmax gating trick is a differentiable sampling method for discrete distributions. It introduces Gumbel noise to create a continuous, differentiable proxy for the otherwise non-differentiable sampling process. For each candidate UGC's representation $\mathbf{x}_r$, we first use a FC layer $f_{\text{score}}(\cdot)$ to output a single scalar logit ℓ_r, which represents the UGC's unnormalized selection probability: $\ell_r = f_{\text{score}}(\mathbf{x}_r)$. To select a subset of size k', we adapt the Gumbel-Softmax trick to perform a differentiable "soft" selection. Given the retrieval set representations as a matrix $\mathbf{R} = [\mathbf{x}_{r_1}, \mathbf{x}_{r_2}, \ldots, \mathbf{x}_{r_k}]^\top \in \mathbb{R}^{k \times d}$ and the corresponding logit vector $L = [\ell_1, \ell_2, \ldots, \ell_k]$, we generate a selection matrix $\mathbf{S} \in \mathbb{R}^{k' \times k}$ through k' independent sampling operations. For the i-th selection (where $i \in \{1, \ldots, k'\}$), a k-dimensional probability vector, which forms the i-th row of $\mathbf{S}$, is generated. Its j-th component is calculated as:

$$\mathbf{S}_{i,j} = \frac{\exp\left(\left(\ell_j + \xi_{i,j}\right)/\tau\right)}{\sum_{l=1}^{k} \exp\left(\left(\ell_l + \xi_{i,l}\right)/\tau\right)}, \tag{7}$$

where $\xi_{i,j} \sim \text{Gumbel}(0,1)$ is an *i.i.d.* noise sample and τ is the temperature parameter that controls the smoothness of the approximation. Finally, we obtain the embedding matrix $\mathbf{R}' \in \mathbb{R}^{k' \times d}$ of the selected UGCs via a matrix multiplication: $\mathbf{R}' = \mathbf{S} \cdot \mathbf{R}$.

Representation Refinement via Information Bottleneck Although the Gumbel-Softmax gate selects a relevant subset, the embeddings in $\mathbf{R}'$ can still contain task-irrelevant noise and redundancy. We therefore introduce a conditional IB that produces query-conditioned, compressed embeddings $\mathbf{R}'_{\text{IB}}$ which are informative for predicting p_i while discarding superfluous variability. Let $\mathbf{R}' = \{\mathbf{x}'_{r_j}\}_{j=1}^{k'}$ and query UGC embedding $\mathbf{x}_{c_i}$, we define a factorized variational posterior:

$$q_\phi\left(\mathbf{Z} \mid \mathbf{R}', \mathbf{x}_{c_i}\right) = \prod_{j=1}^{k'} q_\phi\left(\mathbf{z}_{r_j} \mid \mathbf{x}'_{r_j}, \mathbf{x}_{c_i}\right), \tag{8}$$

$$q_\phi\left(\mathbf{z}_{r_j} \mid \mathbf{x}'_{r_j}, \mathbf{x}_{c_i}\right) = \mathcal{N}\left(\boldsymbol{\mu}_j, \operatorname{diag}\left(\boldsymbol{\sigma}_j^2\right)\right), \tag{9}$$

with parameters

$$\left(\boldsymbol{\mu}_j, \log \boldsymbol{\sigma}_j^2\right) = f_{\text{IB}}\left(\left[\mathbf{x}'_{r_j}; \mathbf{x}_{c_i}\right]\right), \tag{10}$$

and the reparameterization $\mathbf{z}_{r_j} = \boldsymbol{\mu}_j + \boldsymbol{\sigma}_j \odot \boldsymbol{\epsilon}_j$, $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The prior is a standard normal. The overall training minimizes the prediction loss (mean squared error, MSE) and a variational conditional IB loss:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}}\left(\mathcal{M}\left(\mathbf{x}_{c_i}, \mathbf{R}'_{\text{IB}}\right), p_i\right) \tag{11}$$

$$+ \beta \sum_{j=1}^{k'} \text{KL}\left(q_\phi\left(\mathbf{z}_{r_j} | \mathbf{x}'_{r_j}, \mathbf{x}_{c_i} \| \mathcal{N}\left(\mathbf{z}_{r_j} \mid \mathbf{0}, \mathbf{I}\right)\right)\right), \tag{12}$$

where β controls the information-compression trade-off and helps to stabilize the model training. The IB-refined retrieval matrix is:

$$\mathbf{R}'_{\text{IB}} = \left[\mathbf{z}_{r_1}, \mathbf{z}_{r_2}, \ldots, \mathbf{z}_{r_{k'}}\right]^\top \in \mathbb{R}^{k' \times d_z}, \tag{13}$$

which replaces $\mathbf{R}'$ for the aggregation and prediction.

Dataset	ICIP	SMPD	Instagram
# UGCs	20,337	305,613	297,865
# users	17,302	38,312	33,935
avg. popularity	200.78	493.14	4,694.26

Table 1: Dataset Statistics

Robust Prediction With Test-Time Perturbations

To enhance the robustness of our model during inference, we apply test-time perturbations on the joint (image-text) embedding with Gaussian noise at multiple scales and aggregate the resulting predictions with similarity-based weights. For an input embedding $\mathbf{x}$, we generate a set of T perturbed versions $\{\tilde{\mathbf{x}}^{(t)}\}_{t=1}^{T}$ by introducing Gaussian noise. The final prediction $\hat{p}$ is a weighted average of the individual predictions derived from these perturbed embeddings. The weights are based on the cosine similarity between the original and perturbed joint embeddings, giving greater influence to predictions from augmentations that are closer to the original input.

Let $\hat{p}_i^{(t)}$ be the prediction from the perturbed embedding. A confidence score w_t for each prediction is calculated based on its similarity to the original embedding:

$$w_t = 1/\left(1 - \cos\left(\mathbf{x}, \tilde{\mathbf{x}}^{(t)}\right) + \epsilon\right), \tag{14}$$

where $\cos(\cdot, \cdot)$ is the cosine similarity and ϵ is a small constant for numerical stability. The final robust prediction $\hat{p}_i$, is then computed as the normalized weighted average of the individual predictions:

$$\hat{p}_i = \left(\sum_{t=1}^{T} w_t \hat{p}_i^{(t)}\right) / \left(\sum_{t=1}^{T} w_t\right). \tag{15}$$

4 Experiments

We now report results for social media popularity prediction, comparing our proposed model with baselines, covering main experiments, right tail prediction, ablation studies, parameter sensitivity, robustness, and case studies.

Dataset	ICIP			SMPD			Instagram		
Metric	MSE	MAE	SRC	MSE	MAE	SRC	MSE	MAE	SRC
SVR	1.9009	0.8941	0.5241	6.2996	2.0208	0.2163	7.0534	1.9695	0.4035
HyFea	1.9013	1.0181	0.4497	4.7429	1.7080	0.4677	4.7132	1.6924	0.4708
MFTM	1.8970	0.9772	0.4156	4.0222	1.5481	0.5849	4.3073	1.6132	0.5321
CLSTM	1.8724	0.9823	0.4654	3.9143	1.5005	0.5888	4.2431	1.5882	0.5396
HMMVED	1.8556	0.9497	0.4524	3.7154	1.3636	0.6352	4.2461	1.6017	0.5385
DLBA	2.2290	1.0097	0.3614	4.8693	1.7021	0.4387	5.1425	1.7527	0.4007
MASSL	1.9446	0.9278	0.4499	5.5670	1.8427	0.5271	7.8583	2.2274	0.5188
BLIP	2.0646	0.9961	0.3603	4.3884	1.6340	0.5269	5.2436	1.8058	0.3762
CBAN	1.8098	0.9309	0.4727	4.0443	1.5123	0.5754	4.2808	1.5894	0.5426
NIPA	1.9999	0.9980	0.3989	4.2538	1.6532	0.4086	4.0209	1.5565	0.5696
MMRA	1.7600	0.8684	0.5439	3.5119	1.3730	0.6423	3.9456	1.5070	0.5806
SKAPP	0.9662	0.6367	0.6965	1.8196	0.8249	0.8414	2.0936	1.0369	0.8272
JRPP	**0.8972**	**0.6173**	**0.7125**	**1.5178**	**0.7981**	**0.8697**	**1.5627**	**0.8473**	**0.8704**

Table 2: Popularity prediction performance comparison of the baselines and JRPP model on three large-scale UGC datasets.

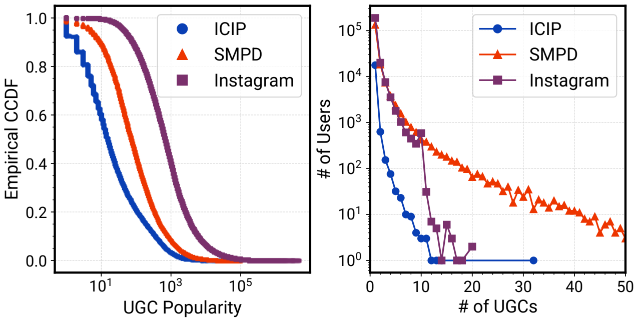

Figure 3: Dataset distributions. Left: Empirical CCDF of UGC popularity. Right: UGCs per user.

Dataset	Metric	SKAPP		JRPP	
		20%	10%	20%	10%
ICIP	MSE	3.175	4.907	**2.648**	**3.712**
	MAE	1.422	1.849	**1.295**	**1.530**
	SRC	0.429	0.231	**0.454**	**0.252**
SMPD	MSE	3.381	5.231	**3.068**	**4.581**
	MAE	1.329	1.683	**1.207**	**1.479**
	SRC	0.471	0.296	**0.508**	**0.372**
Instagram	MSE	3.173	3.900	**2.966**	**3.512**
	MAE	1.464	1.629	**1.358**	**1.496**
	SRC	0.605	0.618	**0.704**	**0.678**

Table 3: Tail prediction performance on the top 20% and top 10% most popular test UGCs.

Experimental Settings

Datasets We evaluate on three large-scale, real-world social media UGC datasets: *ICIP* (Ortis, Farinella, and Battiato 2019), *SMPD* (Wu et al. 2023), and *Instagram* (Kim et al. 2020). Summary statistics and distributional characteristics are provided in Table 1 and Figure 3, respectively.

Baselines We compare our model with 12 strong baselines for predicting social media popularity, including feature-engineering-based: SVR (Khosla, Das Sarma, and Hamid 2014), HyFea (Lai, Zhang, and Zhang 2020), and MFTM (Hsu et al. 2023); deep-learning-based: 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); and retrieval-based: NIPA (Ji et al. 2023a), MMRA (Zhong et al. 2024), and SKAPP (Xu et al. 2025).

Metrics Following previous works (Cappallo, Mensink, and Snoek 2015; Xu et al. 2025; Wu et al. 2023), we use mean squared error (MSE), mean absolute error (MAE), and Spearman's rank correlation (SRC) as evaluation metrics.

Implementation Details We use BLIP (Li et al. 2022) and ViT (Dosovitskiy et al. 2021) as the pre-trained models for extracting text and image embeddings, and the embedding size is set to 768. The split of the three datasets is 8:1:1 for training/validation/test sets. The training set is used as the knowledge database. The popularity is log-transformed. We train our model by the Adam optimizer with a learning rate of 10^{-4}. The batch size is 512, the retrieval number k' after the Gumbel-Softmax gating is $\lfloor 0.8k \rfloor$, the number of low-rank embeddings M in the MoL module is 16, the number of test-time perturbations T is 20. Source code: https://github.com/LOCK233/jrpp

Experimental Results

Performance Comparison The popularity prediction result is reported in Table 2. Across ICIP, SMPD, and Instagram, JRPP consistently outperforms all baselines. Relative to the best counterpart SKAPP, JRPP reduces MSE by 7.14%, 16.59%, and 25.36%, and MAE by 3.05%, 3.25%, and 18.29% on ICIP, SMPD, and Instagram, respectively. It

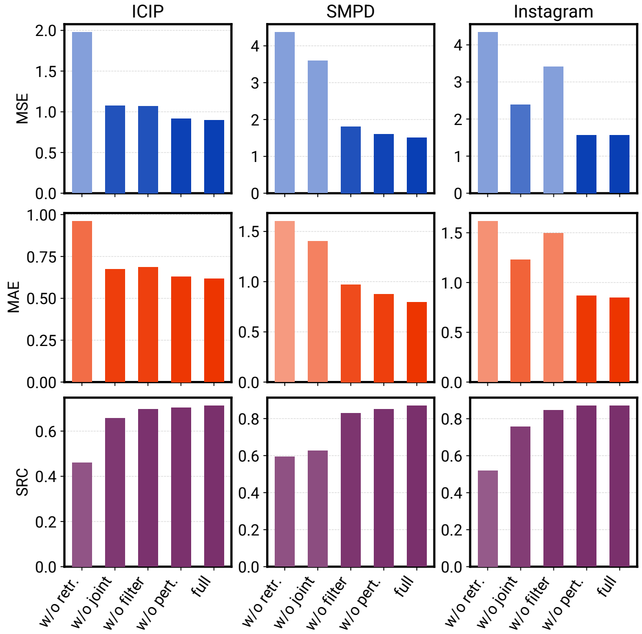

Figure 4: Ablation of the key modules on three datasets.

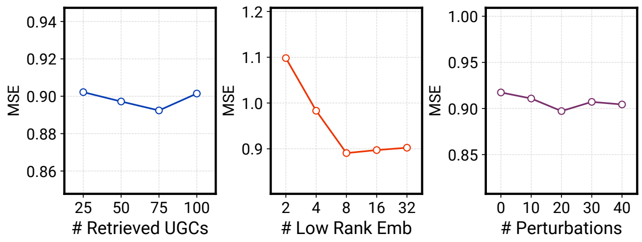

Figure 5: Hyperparameter sensitivity analysis.

also increases SRC by 0.0160 (+2.30%), 0.0283 (+3.36%), and 0.0432 (+5.22%), indicating superior predictive accuracy and ranking quality. The end-to-end joint optimization aligns the retrieval process directly with the prediction objective, while the Mixture-of-Logits-based retrieval module provides more expressiveness. The uncertainty-aware filter prunes noisy and irrelevant UGCs, and a test-time perturbation strategy enhances robustness, yielding more accurate and reliable social media UGC popularity predictions. These results validate the effectiveness of the proposed JRPP framework for jointly retrieving relevant UGCs and predicting future popularity.

Tail Performance To evaluate the performance on highly popular UGCs, we select the top 20% and 10% of the test UGCs based on ground-truth popularity. We compare our model's performance with the best baseline SKAPP. The results are shown in Table 3. We can observe that both models encountered significant performance drops. Nevertheless, our model consistently outperforms SKAPP across all datasets and metrics. This robust outperformance in the tail distribution highlights JRPP's enhanced capability to predict the popularity of popular UGCs.

Ablation Study To validate the contribution of each module in JRPP, we perform ablations on four variants: (1) *w/o retrieval*: this variant completely removes the retrieval module and degenerates into a non-retrieval traditional model; (2) *w/o joint*: this variant replaces the MoL-based joint retrieval and prediction module with a standard dot-product-based retriever; (3) *w/o filter*: this variant removes the uncertainty-aware filtering mechanism; and (4) *w/o perturbation*: this variant omits the test-time perturbations used to obtain a confidence-weighted prediction. The ablation results are shown in Figure 4. Across all three datasets, the full framework attains the lowest MSE/MAE and the highest SRC, indicating that each component of JRPP contributes to the performance; Removing the retrieval causes the largest degradation, underscoring that high-quality, task-aligned retrieval is foundational; Replacing the MoL-based joint module with a standard retriever yields substantial drops across all metrics, confirming the benefits of aligning retrieval directly with the downstream prediction objective rather than a surrogate similarity objective; Removing the uncertainty-aware filter notably increases prediction error, showing the importance of identifying and refining the most relevant UGCs from the initial retrieved set; At last, removing the test-time perturbations produces a small but consistent decline, suggesting that confidence-based ensembling enhances reliability across diverse UGC scenarios.

In-Depth Analysis

Parameter Sensitivity Analysis Here we conduct a sensitivity analysis on three important hyperparameters of our framework: k, the number of retrievals in the joint retrieval and prediction module; M, the number of the low-rank embeddings used in the MoL module; and T, the number of test-time perturbations. The analysis results on the ICIP dataset are shown in Figure 5. We have the following observations: (1) The performance of our model improves as k increases from small values, reflecting that our model is taking the retrieved UGCs as a context augmentation that learns better target UGC representations. Beyond a moderate value, gains saturate and start to decline, which suggests that retrieving too many UGCs may introduce noisy, irrelevant UGCs that harm the prediction; (2) Increasing M enhances the expressivity of the gated similarity (more facets of the content and context alignment) and yields steady gains at low-to-moderate values. However, larger M degrades the prediction performance, which can be attributed to the increased complexity and the risk of overfitting; and (3) Test-time perturbations act as a lightweight ensemble: increasing T initially reduces variance and improves robustness and performance slightly, after which the improvements become insignificant. A large T provides negligible benefit and can degrade performance by averaging over low-confidence perturbations.

Case Study: Retrieval Examples

To illustrate how JRPP's retrieval module operates in practice, we select three query UGCs from the three datasets and examine their retrieval results in Figure 6. In the first case, the query UGC is a photograph of a sunset over a body of

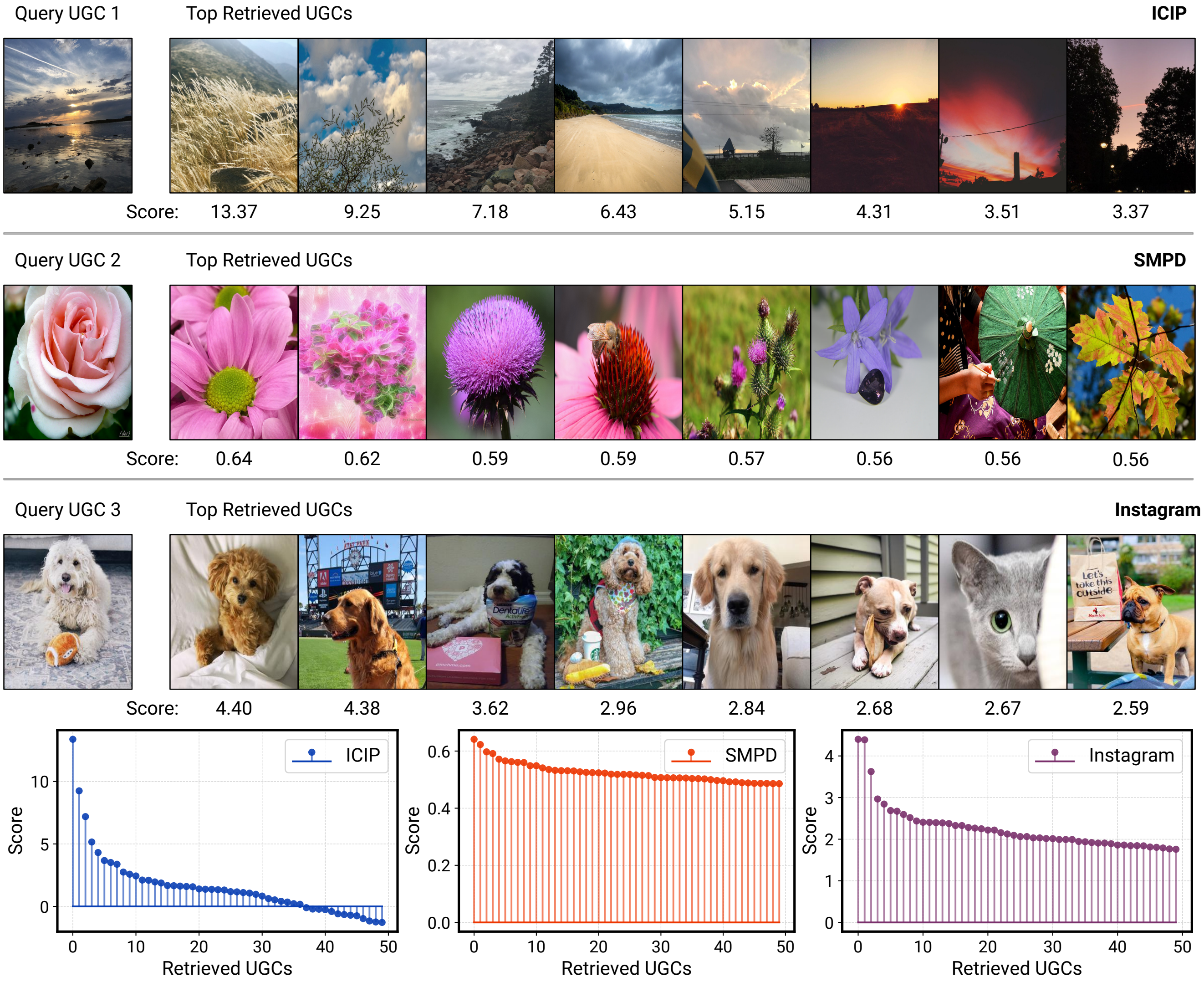

Figure 6: Case study of the top retrieved UGCs given three example UGCs from the three datasets. The UGC photos are presented. Below each photo is a selection score defined in the MoL retrieval module. Bottom: Three selection score distributions for the top-50 retrieved UGCs of the three query UGCs, respectively.

water. Our MoL-based joint retrieval and prediction framework successfully identifies a set of relevant UGCs, prioritizing other landscape and sunset scenes with similar color palettes and compositions. The numbers under the photos are the selection scores defined in the MoL-based retrieval module, which indicates the similarity learned by the model between the query UGC and each retrieved UGC. The second case UGC is a close-up photograph of a pink rose, the retrieval module predominantly returns macro shots of flowers exhibiting similar pink-purple hues and petal structures. In the third case, the query UGC portrays a small dog indoors with a toy, and the retrieved UGCs are largely dominated by companion-animal photographs, particularly dogs. We can also observe that the selection scores decrease as the retrieved UGCs become less relevant to the query UGC—the seventh retrieved UGC is a cat photo.

5 Conclusion

We present JRPP, a joint retrieval and prediction framework for social media UGC popularity that aligns retrieval with the end task and mitigates retrieval noise. JRPP integrates a MoL-based retrieval module for task-aligned matching, an uncertainty-aware filter that couples differentiable subset selection with an information bottleneck refinement, and a confidence-weighted test-time perturbation scheme for robust prediction. Across three large-scale social UGC datasets, JRPP achieves state-of-the-art results on MSE, MAE, and SRC metrics, surpassing strong baselines. These gains demonstrate that tightly coupling retrieval and prediction yields calibrated context integration and improved generalization. Future directions include better retrieval strategies that are capable of acquiring online social knowledge.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant Nos. 62425208, U22A2097 and 62572097.

References

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*, 1–13.

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.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can Cascades e Predicted? In *WWW*, 925–936.

Cheng, Z.; Zhang, J.; Xu, X.; Trajcevski, G.; Zhong, T.; and Zhou, F. 2024. Retrieval-Augmented Hypergraph for Multimodal Social Media Popularity Prediction. In *KDD*, 445–455.

Cheung, T.-h.; and Lam, K.-m. 2022. Crossmodal Bipolar Attention for Multimodal Classification on Social Media. *Neurocomputing*, 514: 1–12.

Ding, B.; and Zhai, J. 2025. Retrieval with Learned Similarities. In *WWW*, 1626–1637.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 1–21.

Drolsbach, C. P.; and Pröllochs, N. 2023. Believability and Harmfulness Shape the Virality of Misleading Social Media Posts. In *WWW*, 4172–4177.

Gao, X.; Jia, X.; Yang, C.; and Chen, G. 2020. Using Survival Theory in Early Pattern Detection for Viral Cascades. *TKDE*, 34(5): 2497–2511.

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.

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 Multimedia*, 9451–9455.

Hsu, C.-C.; Lee, C.-M.; Lin, Y.-F.; Chou, Y.-S.; Jian, C.-Y.; and Tsai, C.-H. 2024. Revisiting Vision-Language Features Adaptation and Inconsistency for Social Media Popularity Prediction. In *ACM Multimedia*, 11464–11469.

Jeon, H.; Lee, J.-e.; Yun, J.; and Kang, U. 2024. Cold-Start Bundle Recommendation via Popularity-based Coalescence and Curriculum Heating. In *WWW*, 3277–3286.

Ji, L.; Park, C. H.; Rao, Z.; and Chen, Q. 2023a. Neural Image Popularity Assessment with Retrieval-augmented Transformer. In *ACM Multimedia*, 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 *KDD*, 930–940.

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.

Krumm, J.; Davies, N.; and Narayanaswami, C. 2008. User-Generated Content. *IEEE Pervasive Computing*, 7(4): 10–11.

Lai, X.; Zhang, Y.; and Zhang, W. 2020. HyFea: Winning Solution to Social Media Popularity Prediction for Multimedia Grand Challenge 2020. In *ACM Multimedia*, 4565–4569.

Lang, J.; Hong, R.; Xu, J.; Li, Y.; Xu, X.; and Zhou, F. 2025. Biting Off More Than You Can Detect: Retrieval-Augmented Multimodal Experts for Short Video Hate Detection. In *WWW*, 2763–2774.

Li, H.; Xia, C.; Wang, T.; Wen, S.; Chen, C.; and Xiang, Y. 2021. Capturing Dynamics of Information Diffusion in SNS: A Survey of Methodology and Techniques. *ACM Computing Surveys*, 55(1): 1–51.

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.

Naab, T. K.; and Sehl, A. 2017. Studies of User-Generated Content: A Systematic Review. *Journalism*, 18(10): 1256–1273.

Ortis, A.; Farinella, G. M.; and Battiato, S. 2019. Prediction of Social Image Popularity Dynamics. In *ICIAP*, 572–582.

Sun, J.; Chen, C.; Hou, C.; Wu, Y.; and Yuan, X. 2025. Multimodal Taylor Series Network for Misinformation Detection. In *WWW*, 2540–2548.

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): 8.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The Information Bottleneck Method. In *Annual Allerton Conference on Communication, Control, and Computing*, 368–377.

Tishby, N.; and Zaslavsky, N. 2015. Deep Learning and the Information Bottleneck Principle. In *IEEE Information Theory Workshop*, 1–5.

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 Multimedia*, 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.

Xie, J.; Zhu, Y.; Zhang, Z.; Peng, J.; Yi, J.; Hu, Y.; Liu, H.; and Chen, Z. 2020. A Multimodal Variational Encoder-decoder Framework for Micro-Video Popularity Prediction. In *WWW*, 2542–2548.

Xu, X.; Zhang, Y.; Zhou, F.; and Song, J. 2025. Improving Multimodal Social Media Popularity Prediction via Selective Retrieval Knowledge Augmentation. In *AAAI*, 932–940.

Xu, X.; Zhou, F.; Zhang, K.; and Liu, S. 2022. CCGL: Contrastive Cascade Graph Learning. *TKDE*, 35(5): 4539–4554.

Xu, X.; Zhou, F.; Zhang, K.; Liu, S.; and Trajcevski, G. 2021. CasFlow: Exploring Hierarchical Structures and Propagation Uncertainty for Cascade Prediction. *TKDE*, 35(4): 3484–3499.

Zhai, J.; Gong, Z.; Wang, Y.; Sun, X.; Yan, Z.; Li, F.; and Liu, X. 2023. Revisiting Neural Retrieval on Accelerators. In *KDD*, 5520–5531.

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 *KDD*, 1513–1522.

Zhong, T.; Lang, J.; Zhang, Y.; Cheng, Z.; Zhang, K.; and Zhou, F. 2024. Predicting Micro-Video Popularity via Multi-Modal 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.

Zhou, M.; Lin, Y.; Liu, G.; Li, Z.; Liao, H.; and Mao, R. 2024. Modeling Personalized Retweeting Behaviors for Multi-Stage Cascade Popularity Prediction. In *IJCAI*, 2598–2606.

Zhu, K.; Feng, X.; Du, X.; Gu, Y.; Yu, W.; Wang, H.; Chen, Q.; Chu, Z.; Chen, J.; and Qin, B. 2024. An Information Bottleneck Perspective for Effective Noise Filtering on Retrieval-Augmented Generation. In *ACL*, 1044–1069.