Modeling Multimodal Information Cascade on Social Media with Interpretable Mixture of Experts

Xin Jing
State Key Laboratory of IOTSC,
University of Macau,
Macau, China
yc27431@um.edu.mo

Zeyu Shi
State Key Laboratory of IOTSC,
University of Macau,
Macau, China
mc46586@um.edu.mo

Zhangtao Cheng
University of Electronic Science and
Technology of China,
Chengdu, China
zhangtao.cheng@outlook.com

Yichen Jing
State Key Laboratory of IOTSC,
University of Macau,
Macau, China
mc35259@um.edu.mo

Yuhuan Lu
State Key Laboratory of IOTSC,
University of Macau,
Macau, China
yc17462@um.edu.mo

Bangchao Deng
State Key Laboratory of IOTSC,
University of Macau,
Macau, China
yc37980@um.edu.mo

Dingqi Yang*
State Key Laboratory of IOTSC,
University of Macau,
Macau, China
dingqiyang@um.edu.mo

Abstract

Information popularity prediction is a crucial yet challenging task for studying the dynamics of information diffusion patterns on social media platforms, which can benefit a wide range of applications, such as misinformation detection and viral marketing. While existing methods have achieved favorable performance in popularity predictions, they primarily focus on modeling the structural and temporal characteristics of information cascades. However, we argue in this paper that exploring other informative signals, such as textual content, can be critical to boost popularity prediction accuracy. Against this background, we propose a novel problem setting: **multimodal information cascade modeling**, which incorporates four essential elements, including information cascade dynamics, user profiles, textual content, and visual content, and we construct the corresponding benchmark datasets to support this problem. Subsequently, we propose **MMCas**, a novel approach for the MultiModal information Cascade popularity prediction task, which is designed to subtly capture the characteristics of the above four elements and, more importantly, their inherent multimodal interactions. Specifically, we first leverage diverse feature extraction pipelines of the four multimodal elements. We then design a Mixture of Experts (MoE) interaction mechanism for the modality fusion and deploy a reweighting module that assigns importance scores for the output of each interaction expert, providing both local and global interpretation. Extensive experiments conducted on our curated datasets demonstrate that MMCas significantly outperforms a wide range of state-of-the-art methods, yielding 3.9%–20.6% improvement over the best-performing baselines across all datasets.

CCS Concepts

• **Information systems → Social networks**.

Keywords

Social networks, multimodal information cascade, popularity prediction, mixture of experts

ACM Reference Format:
Xin Jing, Zeyu Shi, Zhangtao Cheng, Yichen Jing, Yuhuan Lu, Bangchao Deng, and Dingqi Yang. 2026. Modeling Multimodal Information Cascade on Social Media with Interpretable Mixture of Experts. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3774904.3793030

1 Introduction

The rapid spread of information on social media platforms has profound implications for societal well-being, from shaping public opinion to mobilizing collective action for social good. One of the key research problems in this direction is information cascade popularity prediction [2, 25, 38], which aims to forecast the future popularity of given content, like predicting the number of retweets of a given tweet within a given period in the future. Accurately predicting the popularity of information cascades is not only a fundamental challenge in Web research but also a critical building component for various Web applications such as misinformation detection [16], recommendation [28], and risk management [25].

In recent years, numerous studies have focused on modeling the dynamic diffusion process based on information cascades for popularity prediction [2, 6, 32, 41]. Deep-learning-based approaches

[*] Corresponding author.

WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3793030

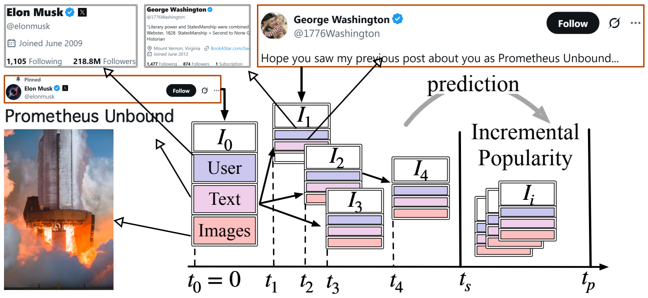

Elon Musk
@elonmusk
Joined June 2009
1,105 Following 218.8M Followers
George Washington
@1776Washington
"Literary power and StatesManship were combined Webster, 1828 StatesManship = Second to None G Historian
Mount Vernon, Virginia BookAStar.com/Ge
Joined June 2012
1,477 Following 874 Followers 1 Subscription
George Washington
@1776Washington
Follow
Hope you saw my previous post about you as Prometheus Unbound...
Pinned
Elon Musk
@elonmusk
Follow
Prometheus Unbound
I_0
User
Text
Images
I_1
I_2
I_3
I_4
I_i
prediction
Incremental Popularity
$t_0 = 0$
t_1
t_2
t_3
t_4
t_s
t_p

Figure 1: A toy example of a multimodal cascade graph; Elon Musk published a tweet I_0 at t_0 with the following retweets $\{I_1, I_2, I_3, I_4\}$ under the observation t_s; The task is to predict the incremental popularity between t_s and t_p.

have become predominant due to their superior prediction performance. These methods typically follow a two-step procedure. First, they employ graph embedding techniques to learn low-dimensional representations of nodes within a cascade, capturing the structural properties of the cascade graph [5, 29, 42]. Second, they model the temporal characteristic of the cascade by sequentially feeding the learned node embeddings into sequence models [3, 35] to capture dependencies among the nodes [5, 32, 42] or modeling continuous-time dynamics of cascades [7, 14, 15] so as to forecast the future diffusion trends.

Although these approaches have achieved favorable performance in popularity predictions, they primarily focus on modeling the structural and temporal characteristics of information cascades. In this paper, *we argue that exploring other informative signals that complement the structural and temporal cascade features is also critical for approaching real-world scenarios and enhancing the accuracy of popularity predictions.* To this end, we incorporate information cascade dynamics, user profiles, textual content, and visual content to further enrich the cascade representations. This integrative approach allows us to capture not only how information spreads but also why it resonates, accounting for factors such as author credibility, emotional tone, and visual appeal, all of which are pivotal in campaigns for health awareness, environmental advocacy, or humanitarian aid. First, user profiles, including verification status and follower/following counts, play a pivotal role in shaping individual information propagation dynamics. Second, multimodal user-generated contents (UGCs), including textual and visual elements, inherently convey semantic, emotional, and contextual cues that influence audience engagement and the propagation process. Figure 1 shows a real-world information cascade diffusion process on Twitter. Elon Musk (User Profile) published a tweet I_0 at time t_0 with the text “Prometheus Unbound” (Textual) and an image of a rocket (Visual). Then George Washington (User Profile) retweeted it at time t_1 and created a new tweet I_1 with the text “Hope you ...” (Textual), establishing a retweet chain (I_0, I_1, t_1); multiple retweet chains formed the cascade dynamics (Cascade). The four fundamental elements, including user profile, textual, visual content, and cascade, collectively impact the future popularity of the original tweet. As summarized in Table 1, some UGC-based works [8, 31, 36, 39] focus on modeling textual and visual content while some works [2, 7, 14, 15, 18, 20, 32, 42] solely rely on the information cascade, however, there is currently no existing method capable of effectively capturing all four aspects simultaneously.

Table 1: Comparison of our proposed MMCas with existing information popularity prediction approaches.

Model	User Profile	Textual	Vision	Cascade
DTCN [31]	✗	✗	✓	✗
UHAN [36]	✗	✓	✓	✗
CBAN [8]	✗	✓	✓	✗
MMRA [39]	✗	✓	✓	✗
DeepCas [18]	✗	✗	✗	✓
DeepHawkes [2]	✗	✗	✗	✓
VaCas [42]	✗	✗	✗	✓
CasFlow [32]	✗	✗	✗	✓
CTCP [20]	✗	✗	✗	✓
ConCat [15]	✗	✗	✗	✓
CasDo [7]	✗	✗	✗	✓
CasFT [14]	✗	✗	✗	✓
MMCas (ours)	✓	✓	✓	✓

Against this background, we address in this study the information popularity prediction problem based on such multimodal cascades, encompassing the four essential elements: user profiles, textual content, visual content, and cascade structure. However, effectively modeling and leveraging multimodal cascades poses several critical challenges: **Challenge 1: How to fuse these heterogeneous modalities is non-trivial**: UGC-based methods [8, 31, 36, 39] leverage some multimodal fusion techniques that are designed for aligned data like image-text pairs and cannot be directly applied to the information diffusion scenario. The primary challenge lies in harmonizing the continuous-time information diffusion dynamics of cascades with the static, content-based nature of images and text. Simple operations like concatenation often result in inadequate interaction modeling, as they fail to capture the complex and evolving relationship between UGCs and the information diffusion process, where naive concatenation yielded suboptimal performance, as evidenced by our experiments later. **Challenge 2: Influence of different modalities on information popularity is highly contextualized:** As illustrated in Figure 1, a post by Elon Musk has a strong “celebrity effect”, where the user profile modality (e.g., follower count) overwhelmingly drives popularity, often overshadowing the content itself. In specific scenarios such as Twitter hashtags, trending topics can drive popularity independently of influential users, highlighting the significance of textual and visual content. The ability to discern when user influence outweighs content virality or how an engaging image can enhance a textual message is essential not only for enhancing predictive precision but also for providing an interpretable information diffusion process.

To address these challenges, we introduce **MMCas**, a novel approach designed for the MultiModal information Cascade popularity prediction task. First, we leverage pre-trained models to generate dense embeddings for vision and textual content, incorporate user profile features, and adopt neural Ordinary Differential Equations (ODEs) [4] to model the continuous-time dynamics of cascades. Second, after obtaining the representations of user profiles, textual content, visual content, and cascade dynamics, we perform multimodal fusion based on Mixture of Experts (MoEs), designed to explicitly model subtle interaction patterns among the modalities. Specifically, we design multiple interaction experts, where each expert specializes in capturing a distinct type of inter-modal relationship, including *uniqueness experts* for information specific to individual modalities, *synergistic experts* for emergent information arising from the combination, and *redundant experts* for shared information across modalities. What's more, to adaptively integrate these expert outputs, we design a reweighting module that assigns dynamic importance scores to each expert based on the input sample, enabling context-aware fusion and providing *both local and global interpretability by revealing the contribution of each interaction type to the final prediction*. Finally, we implement a step-wise training approach, incorporating weakly self-supervised interaction losses and strongly supervised losses to guide and confine all experts in the model. Due to the lack of multimodal cascade datasets in the literature, we *build the first benchmark datasets* which incorporate all four elements for multimodal cascade modeling. These datasets are collected to rigorously validate our research motivation and design choices. They can serve as a foundational resource for advancing the information cascade research community, providing comprehensive and realistic information diffusion benchmarks for future works. In summary, our contributions are as follows:

- We propose a novel problem setting: multimodal cascade modeling, which incorporates four essential elements, including user profiles, textual content, visual content, and cascade for the information popularity prediction task, and construct the first multimodal cascade benchmark datasets, which will be released publicly and serve as a foundational resource for future studies.
- We propose **MMCas**[1], a novel approach that integrates all four elements of multimodal cascades under a unified framework, by leveraging multiple interaction MoEs to explicitly model diverse interaction patterns among the modalities and subsequently designing a reweighting module for assigning dynamic importance scores to each expert, providing both local and global interpretability.
- Extensive experiments are conducted on our curated datasets to evaluate the effectiveness of MMCas. Results demonstrate that MMCas significantly outperforms a wide range of state-of-the-art baselines on cascade information popularity prediction tasks, yielding 3.9%–20.6% improvement over the best-performing baselines across different datasets.

2 Related Work

We first outline the methods related to information cascade popularity prediction, while the work on multimodal learning is described in the Appendix A. Information cascade popularity prediction has been widely studied in the field of social media analysis. Existing methods can be broadly categorized into three groups: feature engineering methods, statistical methods, and deep learning-based methods. **(1) Feature engineering methods** [6, 13, 23] rely on handcrafted cascade-related features, such as cascade graphs and social relationships, to predict popularity using predefined functions. **(2) Statistical methods** [17, 33, 37, 38] model the process of information diffusion as a sequence of events driven by underlying temporal dynamics. In this context, various point process techniques have been adopted, including Poisson process [25, 27] and Hawkes process [2, 37, 38]. **(3) Deep learning-based methods** [5, 7, 20, 32, 40] aim to construct automated frameworks that leverage Graph Neural Networks (GNNs) or Recurrent Neural Networks (RNNs) like Gated Recurrent Unit (GRU) to effectively model both the temporal and structural dynamics of cascades for popularity prediction. Some latest works [7, 14, 15], including CasDo [7] and CasFT [14] model the continuous-time dynamics of the information cascades with neural ODEs. As a result, deep learning-based methods have emerged as the dominant paradigm for addressing the task of cascade popularity prediction.

Despite this, these works overlook the rich multimedia information associated with information items (e.g., user profile data and user-generated content), thus leading to suboptimal prediction performance. To address this issue, we first propose the multimodal information cascade popularity task, which integrates the cascade dynamics, user-centric, and content-centric perspectives.

3 Problem Definition

Suppose a Twitter user, denoted as u_0, posts a tweet I_0 at time $t_0 = 0$ (we set the original tweet time to 0), where $I_0 = \{user, text, images\}$ contains the user profile, textual and visual information in this tweet as shown in Figure 1. Subsequently, other users can engage with this tweet through various actions, like retweeting. Given an observation time t_s, we define the retweet cascade at time t_s as $C(t_s) = \{(I_{k1}, I_{k2}, t_k)\}_{k \in M}$, where the triplet (I_{k1}, I_{k2}, t_k) means there exists a retweeting action at t_k between I_{k1} and I_{k2} and M indicates there are M triplets involved in the diffusion process and $t_k \leq t_s$. Some definitions are outlined as follows: **1) Multimodal Information Cascade:** Given a tweet I and its corresponding retweet cascade $C(t_s)$ observed at time t_s, its multimodal cascade graph $\mathbf{G}(t_s)$ is defined as $\mathbf{G}(t_s) = (\mathcal{V}_c(t_s), \mathcal{E}_c(t_s))$, where $\mathcal{V}_c(t_s)$ is the tweet set of triplets in $C(t_s)$ and $\mathcal{E}_c(t_s)$ is the edge set in the multimodal cascade graph where an edge presents that there exists a retweeting action between two tweet. **2) Global Graph:** Given all the retweet cascades under the observation time, we define the global graph as $\mathcal{G} = (\mathcal{V}_g, \mathcal{E}_g)$, where the edge in $\mathcal{E}_g$ represents the node relationship, such as the follower/following relationship in the social network. **3) Information Cascade Popularity Prediction:** Given the observed multimodal cascade $C(t_s)$ at t_s, we predict the incremental popularity $Y = |C(t_p)| - |C(t_s)|$, where t_s is the observation time, t_p is the prediction time and $|C|$ denotes the number of triplets in the cascade C.

4 MMCas

In this section, we present the details of our proposed method **MMCas**, as shown in Figure 2, for the MultiModal information Cascade popularity prediction task, which jointly models cascade dynamics, user profiles, textual and vision content.

4.1 Multimodal Feature Extraction

For each given textual content x_i of a tweet I_i, we first tokenize the text, feed the tokenized sequence to a multilingual Bert-based model [9], and use the final hidden state as the textual feature

[1] https://github.com/UM-Data-Intelligence-Lab/MMCas

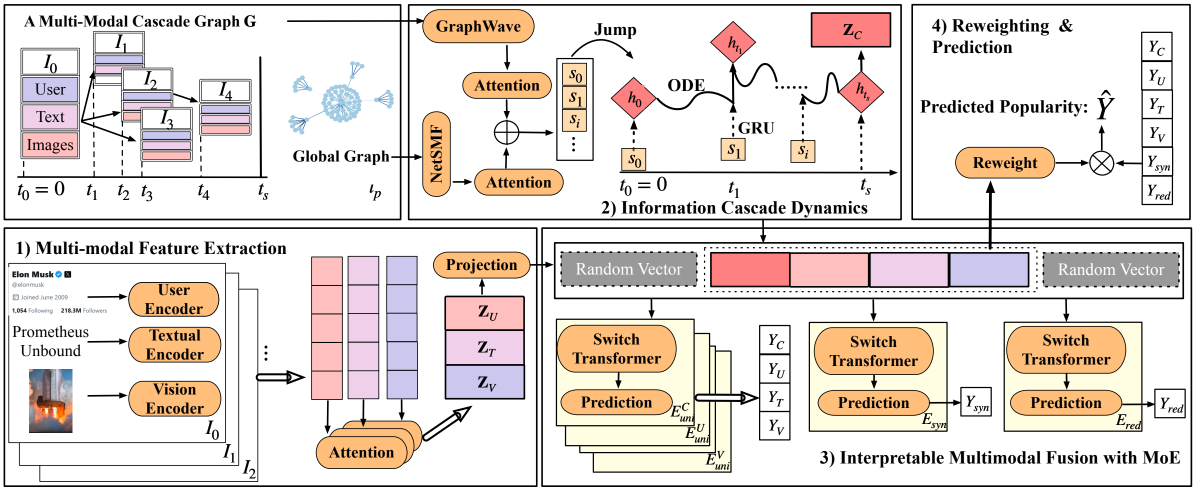


Figure 2: An overview of our proposed model MMCas. (1) MMCas employ different encoder methods for user profile, textual and vision content, respectively, getting corresponding dense embeddings Z_U, Z_T, Z_V. (2) MMCas leverage graph embedding techniques and neural ODEs to capture continuous-time dynamics, getting Z_C. (3) Three types of experts, including uniqueness expert E_{unic}, synergy expert E_{syn} and redundancy expert E_{red} are designed to facilitate the fusion and interaction between the four modalities Z_C, Z_U, Z_T, Z_V. (4) MMCas adopts a reweighting strategy to dynamically combine expert outputs via learned weights for final popularity prediction $\hat{Y}$.

vector Z_{T_i}. Then we integrate all the textual feature vectors Z_{T_i} in the cascade and use an attention module to capture the global dependency, getting the textual representation Z_T for the textual contents in the information cascade.

For the vision content, we leverage another pretrained model, Vision Transformer (ViT) [11] to obtain the visual feature Z_{T_j}. Similar to textual content, we also integrate all the vision feature vectors Z_{Vj} within the cascade and employ another attention module to derive the vision representation Z_V.

For the user profile, we use two functions to represent each part of the user attributes and concatenate them together as the user profile representation:

$$Z_{u_i} = [f_{u_1}(followers, followings) \oplus f_{u_2}(type)], \tag{1}$$

where followers and followings represent the number of followers and followings for user u_i, type indicates the user's verification status (verified/unverified), f_{u_1} is an MLP, and f_{u_2} is a word embedding vector. Each retweet corresponds to a new user, and attention modules are used to model the sequential user patterns, getting the user profile representation Z_U.

4.2 Information Cascade Dynamics

4.2.1 Structural Learning. We employ existing graph embedding methods to model the structure of both the cascade graph and the global graph. For the cascade graph, following previous work [32, 42], we use GraphWave [10] to capture the local structural information. Specifically, given the observed cascade graph $\mathbf{G}(t_s)$ at observation time t_s, we set the weight between nodes as the time interval of retweet triplet time to observation time and leverage heat wavelet diffusion patterns to get each node's low-dimensional embeddings $e_c(u_i)$ in the cascade graph, where $u_i \in \mathcal{V}_c$.

For the structural features of the global graph, due to the large scale of the global graph, we choose the fast and scalable network embedding method NetSMF [24] to get each node's representation $e_g(u_i)$in the global graph, where $\mathcal{G} = (\mathcal{V}_g, \mathcal{E}_g)$ and $u_i \in \mathcal{V}_g$.

4.2.2 Modeling Sequential Information with Self-Attention. After getting the embeddings $E_c(u)$ of each node in the cascade graph and $E_g(u)$ of each node in the global graph without real-valued timestamps, we model the sequential information of the two graphs with two attention modules, getting the hidden representation $s_{c0}, s_{c1}, ..., s_{cN}$ and $s_{g0}, s_{g1}, ..., s_{gN}$ for cascade graph and global graph respectively. Then we concatenate the two hidden representations, denoted as $s_0, s_1, ..., s_N$, and take them as the jump condition in the dynamic flow.

4.2.3 Modeling Continuous Dynamics with Neural ODEs. After getting the jump condition $s_0, s_1, ..., s_N$, we leverage neural ODEs to model the dynamics with a vector representation h_{t_i} at every timestamp t_i that acts as both a summary of the past history and as a predictor of future dynamics. Meanwhile, by making instantaneous updates $s_0, s_1, ..., s_N$ to the hidden state $\mathbf{h}_t$, we can incorporate abrupt changes according to new observed events.

Here we use a standard multi-layer fully connected neural network f_1 to model the continuous change in the form of an ODE. When a new retweeting action occurs at time t_i, we use a GRU function g to model instantaneous changes based on a newly observed node:

$$\frac{dh_{t_i}}{dt} = f_1(t, h_{t_{i-1}}), \tag{2}$$

$$h'_{t_i} = \text{ODESolve}(f_1, h_{t_{i-1}}, (t_{i-1}, t_i)), \tag{3}$$

$$h_{t_i} = g(h'_{t_i}, s_i). \tag{4}$$

By solving Eq. 3, we can get a sequence of hidden states after each jump and we use another ODE to propagate the last event of different cascades to the fixed observation time, where we take the hidden representation h_{t_s} exactly in the observation time as the cascade dynamics representation Z_C.

4.3 Interpretable Multimodal Fusion with MoE

Based on the latent embeddings of the four modalities, including Z_C, Z_U, Z_T, Z_V, the key challenge for the multimodal information cascade popularity prediction is how to capture the subtle interactions between different modalities for efficient fusion. The simplest way is to concatenate all the embeddings of the four modalities, followed by an MLP layer to make the prediction (one of our ablated variants in our experiments later). However, this simple approach does not explicitly account for the heterogeneous interactions (e.g., synergy or redundancy) between information across different modalities, thus resulting in suboptimal performance, as evidenced by our experiments.

Inspired by Partial Information Decomposition (PID) [30] which provides a theoretical framework for comprehending the interactions between modalities and categorizes information into three types: uniqueness for each modality (information specific to each modality), synergy (emergent information resulting from the fusion of all modalities), and redundancy (shared information across all modalities), we design three types of MoEs, including six experts in total, to model the multimodal interaction and fusion: 1) four experts are for unique information of text, vision, user, and cascade dynamics, respectively; 2) one expert is for synergy between the above four modalities, and 3) another expert is for redundancy between these modalities.

4.3.1 Multimodal Interaction and Fusion. Given the four embeddings $Z_C \in \mathbb{R}^{d_c}$, $Z_U \in \mathbb{R}^{d_u}$, $Z_T \in \mathbb{R}^{d_t}$, $Z_V \in \mathbb{R}^{d_v}$ of the four modalities, we first project them into a unified latent space with dimension d using modality-specific linear layers:

$$Z'_M = W_M Z_M + b_M, \text{ for } M \in \{C, U, T, V\} \tag{5}$$

where W_M and b_M are learnable parameters. This projection ensures that all modalities reside in a commensurate vector space, facilitating effective interaction modeling. The fused multimodal representation Z_F is then constructed by concatenating these aligned embeddings:

$$Z_F = [Z'_C, Z'_U, Z'_T, Z'_V]. \tag{6}$$

To capture the rich spectrum of interactions within Z_F, we employ a Mixture of Experts (MoE) architecture comprising six experts. Each expert is a neural network. Here, we use Switch Transformer [12] as our backbone, which specializes in modeling a specific type of modality interaction:

- **Four Uniqueness Experts** (E^C_{uni}, E^U_{uni}, E^T_{uni}, E^V_{uni}): Each expert is dedicated to capturing the information that is unique to one specific modality, including cascade, user, text, and vision.
- **One Synergy Expert** (E_{syn}): This expert focuses on the emergent information that arises only from the non-linear combination of all four modalities.
- **One Redundancy Expert** (E_{red}): This expert models the information that is shared or common across two or more modalities.

Each expert E_k takes the fused representation Z_F as input and outputs a prediction for the popularity:

$$Y_k = E_K(Z_F). \tag{7}$$

4.3.2 Weakly Self-Supervised Expert Specialization via Input Perturbation. To train the experts to specialize in their respective interaction types, we adopt a weakly-supervised strategy guided by principles from PID.

The core idea is to systematically perturb the multimodal input Z_F by replacing one modality's embedding with a random vector $r \sim \mathcal{N}(0, I)$ sampled from a standard normal distribution, thereby ablating information from that modality while preserving others:

$$Z_{F_i} = [Z'_C, ..., r_i, ..., Z'_V], \text{ for } i \in \{C, U, T, V\}, \tag{8}$$

where the $i - th$ position corresponds to one of the four modalities. This perturbation creates four input variants, each missing one modality. By comparing the experts' outputs on the original input Z_F and the perturbed variants Z_{F_i}, we can define specialized losses that enforce distinct interaction behaviors:

Here, Y_k is the output of an expert for the original input while $Y_{ki} = E_k(Z_{F_i})$ is the output of the $i - th$ perturbed input. We define specialized interaction losses for each expert type to encourage the desired behavior. For the uniqueness expert, taking the expert E^C_{uni} as an example, we treat the Z_{F_C} as the negative samples and the left perturbed Z_{F_i} as the positive samples, getting the negative objective Y_{CC} and three positive objective Y_{CU}, Y_{CT}, Y_{CV}. Besides, we regard the original output Y_C as the truth and use the InfoNCE [22] loss as the self-supervised loss function to estimate the mutual information between these outputs and get the loss $\mathcal{L}^{int}_C$. For the synergy expert, we take all the perturbed outputs as negative objectives and encourage Y_{syn} to be dissimilar to all outputs generated from incomplete modality sets, capturing information that requires the full context. For the redundancy expert, take all the perturbed outputs as positive objectives and want Y_{red} to be similar to all outputs from partial inputs, modeling the information that is resilient to the loss of any single modality. Here, we use mean squared logarithmic error (MSLE) to balance the synergy and redundancy expert and get the $\mathcal{L}^{int}_S$, $\mathcal{L}^{int}_R$ loss, respectively.

4.4 Reweighting and Prediction

To ensure each expert k produces meaningful predictions, we impose the Mean Squared Logarithmic Error (MSLE) loss on every expert's output:

$$\mathcal{L}^k_{expert} = \Big(\log(\hat{Y}_k + 1) - \log(Y + 1)\Big)^2, \quad \forall k, \tag{9}$$

where Y is the ground truth popularity.

What's more, to adaptively combine the predictions from the six experts, we employ a reweighting network f, implemented as an MLP, which takes the fused representation Z_F as input and outputs a set of non-negative weights:

$$\mathbf{w} = (w_C, w_U, w_T, w_V, w_{syn}, w_{red}) = \text{softmax}(f(\mathbf{Z}_F)) \tag{10}$$

The final popularity prediction is the weighted sum of the expert outputs:

$$\hat{Y} = \sum_k w_k \cdot Y_k. \tag{11}$$

Table 2: Dataset statistics

Dataset	Twitter	Twitter-Hashtag	Weibo
Cascades	7,205	91,648	18,000
Verified users	1,184	18,360	32,839
Unverified users	225,039	2,530,647	443,151
Avg. followees	575	752	625
Text across cascade event	10.13%	98.9%	64.9%
Vision across cascade	46.66%	78.24%	–
Avg. popularity	31.9	91.9	58.3

The complete training objective combines strong popularity prediction loss, medium task supervision, and weak interaction losses:

$$\mathcal{L} = (log_2(Y+1) - log_2(\hat{Y}+1))^2 + \frac{\lambda_1}{K}\sum \mathcal{L}_k^{int} + \frac{\lambda_2}{K}\sum_k \mathcal{L}_{expert}^k, \tag{12}$$

where Y is the ground truth incremental popularity, $\hat{Y}$ is the predicted incremental popularity, K is the number of experts, λ_1 and λ_2 are hyperparameters. This joint optimization ensures that the model not only makes accurate predictions but also learns interpretable, specialized experts that reflect the underlying multimodal interactions.

5 Experiments

5.1 Datasets

We conduct experiments on three real-world datasets: **Twitter**, **Twitter-Hashtag**, and Sina **Weibo**. Twitter and Twitter-Hashtag datasets are the first multimodal cascade datasets curated by us, encompassing cascade dynamics, user profiles, textual, and visual content. We collect the Twitter and Twitter-Hashtag datasets via the general Twitter Streaming API[2], using the "Spritzer" access level, which is the lightest and shallowest stream and contains approximately 1% of all public tweets. In addition, the Weibo dataset we used in this paper includes cascade, user, and textual content. The details of the data collection process are provided in Appendix B.

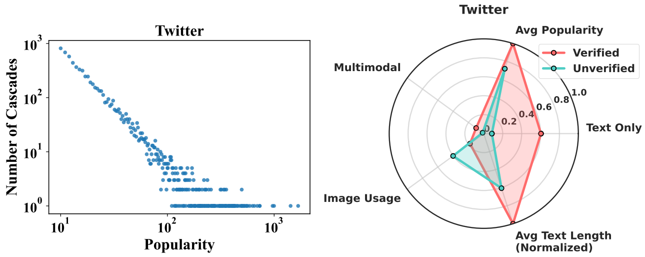


Figure 3: The distribution of popularity (left) and tweet difference between verified and unverified users (right).

5.1.1 Data Analysis. The statistics of the three datasets are summarized in Table 2, which mainly presents the number of cascades, user attributes, text, and vision content percentage in the three datasets. The popularity distributions in detail of all datasets follow a power-law pattern, consistent with observations in previous studies [2, 5, 7, 32], as shown in Figure 3, which also reveals the preferences and influences of verified users and unverified users in terms of tweet posting. Verified users tend to post tweets that incorporate multimodal content and longer texts, suggesting higher quality and greater potential for popularity. More analysis across different datasets is discussed in Appendix B.

5.2 Experimental Settings

5.2.1 Pre-processing. Following previous works [7, 32], we set the observation time t_s of these datasets to the early diffusion proportion, where the normalized popularity propagation process is shown in Figure 3, specifically 1 hour and 3 hours for Twitter, 1 day and 3 days for Twitter-Hashtag, 0.5 hours and 1 hour for Weibo. Accordingly, we set the prediction time t_p to 3 days for Twitter, 28 days for Twitter-Hashtag, and 24 hours for Weibo. For Twitter-Hashtag, we further filter out cascades with fewer than 10 participants within observation time, and we consider the first 100 triplets for cascades with more than 100 triplets, like previous work [20, 32, 42]. For all datasets, we split 70% of the corresponding data for training and the rest for testing (15%) and validation (15%). For the global graph, we construct it by the retweeting relationship like [32].

5.2.2 Baselines. We select ten state-of-the-art baselines of two categories as follows: 1) UGC-based popularity prediction: DTCN [31], UHAN [36], CBAN [8], MMRA [39]; 2) Cascade-based popularity prediction: Feature-based approaches, DeepHawkes [2], CasCN [5], CasFlow [32], CTCP [20], CasDo [7], CasFT [14]. Details of baselines are shown in Appendix C.

5.2.3 Metrics. Following prior work [7, 32], we adopt two widely used evaluation metrics to assess the effectiveness of the proposed MMCas, including mean squared logarithmic error (MSLE) and mean absolute percentage error (MAPE) for evaluation.

5.3 Overall Performance

The experiment results on three datasets comparing baselines and MMCas are shown in Table 3. Our MMCas consistently and significantly outperforms all baselines on three datasets under all evaluation metrics. Compared to the best-performing baselines, MMCas achieves 10.2%-20.6% improvement on MSLE and 3.9%-19% improvement on MAPE with two different settings. These experimental results demonstrate the significance of the multimodal information cascade modeling by MMCas. Specifically, MMCas integrates spatiotemporal patterns of cascades, textual and vision features, and user profiles into a unified framework, achieving more accurate prediction of information popularity.

5.4 Ablation Study & Variants

We have conducted a series of experiments for the ablation study to investigate the contribution of each key component of our MMCas as follows:

[2] https://docs.x.com/home

Table 3: Performance comparison between baselines, MMCas_variants, and MMCas on three datasets under two observation times measured by MSLE, MAPE (lower is better), and the improvement of MMCas over the best-performing baselines. "-" indicates that the corresponding method is designed mainly to process UGCs with visual components.

Method	**Twitter**				**Twitter-Hashtag**				**Weibo**			
	1 hour		**3 hours**		**1 Day**		**3 Days**		**0.5 hours**		**1 hour**	
	MSLE	MAPE	MSLE	MAPE	MSLE	MAPE	MSLE	MAPE	MSLE	MAPE	MSLE	MAPE
DTCN	2.8336	0.4659	2.2064	0.3437	6.8663	0.7534	5.8841	0.5641	-	-	-	-
UHAN	2.1872	0.3281	1.9420	0.2893	6.7598	0.6990	5.7441	0.5088	3.1446	0.4897	2.9741	0.4819
CBAN	2.2364	0.3396	2.0284	0.2913	6.7779	0.7124	5.8314	0.5125	3.1649	0.4911	3.0124	0.2887
MMRA	1.8946	0.3117	1.7915	0.3229	6.6451	0.6954	5.3414	0.5018	-	-	-	-
Feature-based	2.2395	0.3894	2.5959	0.4226	8.5018	0.7466	6.2557	0.5836	3.4561	0.5176	3.2904	0.5146
DeepHawkes	1.8247	0.3012	1.7894	0.3244	6.9714	0.7011	5.9711	0.5164	2.9104	0.4217	2.8714	0.4167
CasCN	1.7641	0.2940	1.7146	0.3054	6.8417	0.6844	5.8147	0.5014	2.8647	0.4187	2.7641	0.4018
CasFlow	1.5078	0.2563	1.4812	0.2716	6.0759	0.6590	4.3101	0.4922	2.5179	0.3735	2.4346	0.3594
CTCP	1.5247	0.2397	1.5116	0.2704	6.2115	0.6124	4.4314	0.4887	2.7140	0.3487	2.4971	0.3411
CasDo	1.4363	0.2448	1.4107	0.2669	5.7045	0.5844	4.1381	0.4701	2.4858	0.3535	2.2904	0.3314
CasFT	1.4058	0.2495	1.4322	0.2601	5.6593	0.6222	4.0419	0.4695	2.4736	0.3552	2.2678	0.3608
MMCas-w/o C	1.5974	0.3047	1.5341	0.3142	5.7419	0.6051	5.1264	0.5013	2.9415	0.4276	2.8795	0.4190
MMCas-w/o U	1.3686	0.2495	1.3315	0.2457	4.6072	0.5074	3.6290	0.4204	2.2260	0.3428	2.0021	0.3345
MMCas-w/o T	1.3564	0.2561	1.3147	0.2355	4.6447	0.4913	3.9148	0.4697	2.2911	0.3449	2.1543	0.3316
MMCas-w/o V	1.2846	0.2244	1.2897	0.2122	4.5250	0.4897	3.5641	0.4167	-	-	-	-
MMCas-concat	1.3765	0.2489	1.3671	0.2597	4.9164	0.5224	3.9345	0.4681	2.3751	0.3524	2.2041	0.3469
MMCas-trans	1.3749	0.2481	1.3619	0.2564	4.8971	0.5207	3.8764	0.4617	2.3347	0.3501	2.1989	0.3441
MMCas-swiT	1.3347	0.2418	1.3318	0.2533	4.8114	0.5157	3.7106	0.4244	2.2874	0.3489	2.1052	0.3347
MMCas	**1.2631**	**0.2231**	**1.2512**	**0.2105**	**4.4915**	**0.4812**	**3.4976**	**0.4074**	**2.0065**	**0.3351**	**1.8015**	**0.3186**
(improvement)	↑ 10.2%	↑ 6.9%	↑ 10.7%	↑ 19.0%	↑ 20.6%	↑ 17.7%	↑ 13.5%	↑ 13.2%	↑ 18.9%	↑ 3.9%	↑ 20.5%	↑ 3.9%

- **MMCas-w/o C & MMCas-w/o U & MMCas-w/o textT & MMCas-w/o V**: we removed cascade dynamics (C), user information (U), text content (T), and visual information (V) correspondingly to explore the impact of each modality on performance.
- **MMCas-concat & MMCas-trans & MMCas-swiT**: We eliminated the multimodal fusion and reweight modules from mmcas. Instead, given the embeddings Z_C, Z_U, Z_T, Z_V of the four modalities, we use simple concatenation (**MMCas-concat**), transformer (**MMCas-trans**), switch transformers [12] (**MMCas-swiT**) as an alternative, following an MLP layer for the popularity prediction.
- **MMCas-CU & MMCas-CT & MMCas-CV & MMCas-UT & MMCas-UV & MMCas-TV**: We also explored the impact of any combination of two modes on popularity prediction, for example, MMCas-CU represents the combination of cascade dynamics (C) and user profiles (U).

The results and comparison of these variants are shown in Table 3. First, compared to MMCAs-w/o C, MMCAs-w/o U, MMCAs-w/o T, and MMCAs-w/o V, MMCas shows 0.74%-37.43% improvement on MSLE across all datasets, demonstrating the importance of modeling cascade dynamics, user profiles, textual and vision content. Second, compared to MMCAs-concat, MMCAs-trans, and MMCAs-swiT, MMCas shows a significant improvement of 5.36%-18.26% on MSLE across all datasets, showing the effectiveness of multimodal fusion with MoE. Furthermore, through a comparative analysis in the ablation study between the three modalities in Table 3 and the two modalities in Table 4, we observe the significant impact of cascade dynamics on information propagation. This deduction, drawn from the ablation study, aligns with the results in the baseline experiments, where cascade-based methods outperform UGC-based methods.

Table 4: Combination of two modalities on Twitter (1 hour), Twitter-Hashtag (1 day), and Weibo (0.5 hours).

Method	Twitter		Twitter-Hashtag		Weibo	
	MSLE	MAPE	MSLE	MAPE	MSLE	MAPE
MMCas-CU	1.3611	0.2566	4.7845	0.5102	2.2911	0.3449
MMCas-CT	1.3751	0.2514	4.7714	0.5067	2.2260	0.3428
MMCas-CV	1.3810	0.2566	4.7919	0.5244	-	-
MMCas-UT	1.6241	0.3174	5.8116	0.6134	2.9415	0.4276
MMCas-UV	1.6289	0.3213	5.9179	0.6197	-	-
MMCas-TV	1.7182	0.3512	5.9824	0.6203	-	-
MMCas	**1.2631**	**0.2231**	**4.4915**	**0.4812**	**2.0065**	**0.3351**

5.5 Parameter Sensitivity

Here, we delve into the influence of λ_1 and λ_2 in the loss function on the model's performance. Illustrated in Figure 4, it is evident that increasing λ_1 results in a degradation of model performance. This highlights the significance of assigning a relatively strong constraint on each expert's output. While λ_2 exhibits a relatively smooth change, a higher λ_2 value may compromise the constraint effect on the experts, leading to a decline in model performance. Consequently, we assign a relatively minor weight to λ_2, establishing a step-wise training objective. Besides, we investigate the influence of hyperparameters, including the modality hidden dimension after projection in Appendix D.

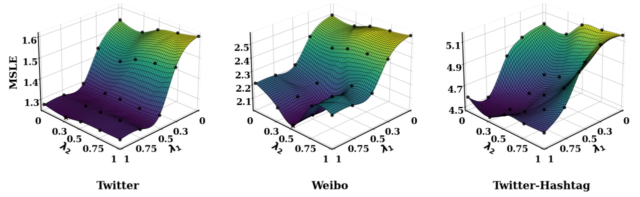


Figure 4: Influence of the hyperparameter in the loss function

5.6 Global Interpretability

We analyze the weights allocated by the reweighting module to each interaction expert across all test samples in Twitter (1 hour) and Twitter-Hashtag. Figure 5 shows the variation in weights across different datasets, providing insights into dataset-level interaction patterns. We see that the reweighting module exhibits the capability to flexibly assign different weights to interaction experts, showcasing its adaptability to capture dataset-specific characteristics. First, the visual expert in Twitter receives less weight than Twitter-Hashtag, because the visual content is more often in Twitter-Hashtag than in Twitter (78.24% v.s. 46.66%). More interestingly, the textual expert in Twitter receives weights with much lower variation than in Twitter-Hashtag, probably because of the different collection mechanisms of the two datasets, where the textual semantics of Twitter-Hashtag are more diverse than Twitter. Specifically, a cascade in Twitter is collected by crawling its retweets, whose topics are usually closely relevant to the original tweet; in contrast, a cascade in Twitter-Hashtag is collected by crawling tweets of the same hashtag, whose topics are more semantically diverse than retweets of a specific tweet.

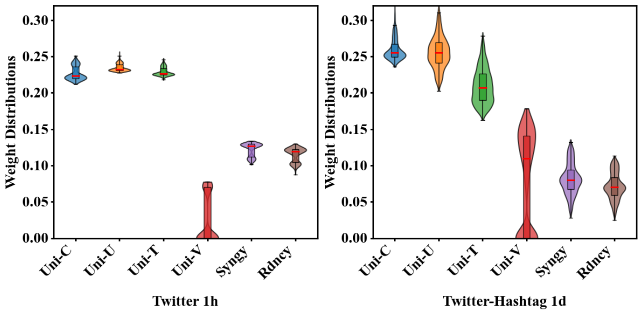
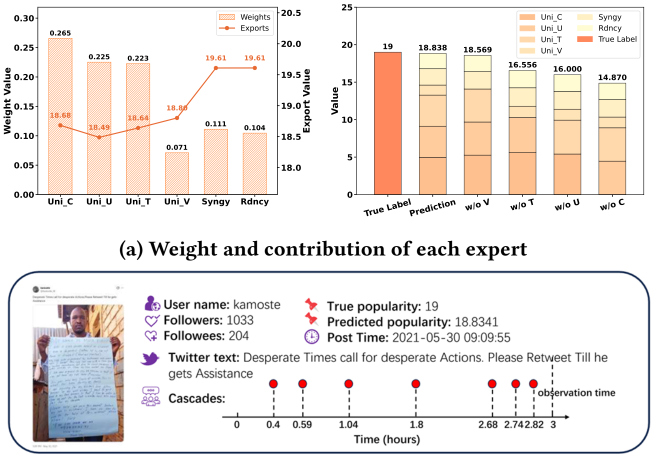


Figure 5: Weight distributions across all modalities on two datasets

5.7 Case Study on Social Good

Figure 6 depicts a scenario from the Twitter dataset where a vulnerable group seeks assistance. It comprises the original tweet content, retweet timestamps, as well as the weights and predicted values of each expert for this instance. An ablation study conducted on this example highlights the significance of cascade dynamics, while indicating that image information holds lesser weight, potentially due to the dense text within the image. This analysis showcases the model's capability to offer **local interpretability** at the sample level. Additionally, we present another case from the Twitter dataset in Appendix E, where the weight assigned to the user profile is highest due to its extensive followees, indicating that MMCas can effectively handle the significance relationship across diverse modalities at a sample-level.



(a) Weight and contribution of each expert

(b) A multimodal information cascade on vulnerable populations

Figure 6: Case study on social good.

6 Conclusion

In this paper, we introduce the problem of multimodal information cascade popularity prediction and release multimodal cascade datasets to support research in this area. To support this task, we propose **MMCas**, a novel framework tailored for the MultiModal Cascades information popularity prediction, effectively capturing both the individual characteristics and the complex interactions among multiple modalities through MoEs. MMCas incorporates specialized experts for uniqueness, synergy, and redundancy, along with a reweighting module that provides both local and global interpretability. Extensive experiments show that MMCas outperforms state-of-the-art baselines, achieving 3.9%–20.6% improvement over the best-performing models. In addition, a case study on vulnerable-group information propagation highlights the potential of MMCas to support social good. In future work, we plan to incorporate multimodal large language models to further improve modality fusion and advance the analysis of multimodal information evolution.

Acknowledgments

This project has received funding from the University of Macau (MYRG-GRG2025-00157-IOTSC), the Science and Technology Development Fund, Macau SAR (0011/2025/RIB1, 001/2024/SKL), and Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC). This work was performed in part at SICC which is supported by SKL-IOTSC, University of Macau.

References

[1] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *TPAMI* 41, 2 (2019), 423–443.

[2] Qi Cao, Huawei Shen, Keting Cen, Wentao Ouyang, and Xueqi Cheng. 2017. Deephawkes: Bridging the gap between prediction and understanding of information cascades. In *CIKM*. 1149–1158.

[3] Guandan Chen, Qingchao Kong, Nan Xu, and Wenji Mao. 2019. NPP: A neural popularity prediction model for social media content. *Neurocomputing* 333 (2019), 221–230.

[4] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. Neural ordinary differential equations. *Advances in neural information processing systems* 31 (2018).

[5] Xueqin Chen, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Fengli Zhang. 2019. Information Diffusion Prediction via Recurrent Cascades Convolution. In *ICDE*. IEEE, 770–781.

[6] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *Proceedings of the 23rd international conference on World wide web*. 925–936.

[7] Zhangtao Cheng, Fan Zhou, Xovee Xu, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Philip S Yu. 2024. Information Cascade Popularity Prediction via Probabilistic Diffusion. *IEEE Transactions on Knowledge and Data Engineering* (2024).

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

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[10] Claire Donnat, Marinka Zitnik, David Hallac, and Jure Leskovec. 2018. Learning structural node embeddings via diffusion wavelets. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1320–1329.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[12] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[13] Shuai Gao, Jun Ma, and Zhumin Chen. 2014. Effective and effortless features for popularity prediction in microblogging network. In *WWW*. 269–270.

[14] Xin Jing, Yichen Jing, Yuhuan Lu, Bangchao Deng, Xueqin Chen, and Dingqi Yang. 2025. CasFT: Future Trend Modeling for Information Popularity Prediction with Dynamic Cues-Driven Diffusion Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11906–11914.

[15] Xin Jing, Yichen Jing, Yuhuan Lu, Bangchao Deng, Sikun Yang, and Dingqi Yang. 2025. On Your Mark, Get Set, Predict! Modeling Continuous-Time Dynamics of Cascades for Information Popularity Prediction. *IEEE Transactions on Knowledge and Data Engineering* (2025).

[16] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[17] Jong Gun Lee, Sue Moon, and Kavé Salamatian. 2012. Modeling and predicting the popularity of online contents with Cox proportional hazard regression model. *Neurocomputing* 76, 1 (2012), 134–145.

[18] Cheng Li, Jiaqi Ma, Xiaoxiao Guo, and Qiaozhu Mei. 2017. Deepcas: An end-to-end predictor of information cascades. In *WWW*. 577–586.

[19] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2247–2256.

[20] Xiaodong Lu, Shuo Ji, Le Yu, Leilei Sun, Bowen Du, and Tongyu Zhu. 2023. Continuous-Time Graph Learning for Cascade Popularity Prediction. *arXiv preprint arXiv:2306.03756* (2023).

[21] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems* 35 (2022), 9564–9576.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[23] Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. 2013. Using early view patterns to predict the popularity of youtube videos. In *WSDM*. 365–374.

[24] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. 2018. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 459–467.

[25] Huawei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. 2014. Modeling and predicting popularity dynamics via reinforced poisson processes. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 28.

[26] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, Vol. 2019. 6558.

[27] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.

[28] Ruijie Wang, Zijie Huang, Shengzhong Liu, Huajie Shao, Dongxin Liu, Jinyang Li, Tianshi Wang, Dachun Sun, Shuochao Yao, and Tarek Abdelzaher. 2021. Dydiffvae: A dynamic variational framework for information diffusion prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 163–172.

[29] Yansong Wang, Xiaomeng Wang, Yijun Ran, Radosław Michalski, and Tao Jia. 2022. CasSeqGCN: Combining network structure and temporal sequence to predict information cascades. *Expert Systems with Applications* 206 (2022), 117693.

[30] Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. 2023. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *Journal of Machine Learning Research* 24, 131 (2023), 1–44.

[31] 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. *arXiv preprint arXiv:1712.04443* (2017).

[32] 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* (2021).

[33] Shuang-Hong Yang and Hongyuan Zha. 2013. Mixture of mutually exciting processes for viral diffusion. In *International Conference on Machine Learning*. PMLR, 1–9.

[34] Haofei Yu, Zhengyang Qi, Lawrence Keunho Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. 2024. Mmoe: Enhancing multimodal models with mixtures of multimodal interaction experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 10006–10030.

[35] Liu Yu, Xovee Xu, Goce Trajcevski, and Fan Zhou. 2022. Transformer-enhanced Hawkes process with decoupling training for information cascade prediction. *Knowledge-Based Systems* 255 (2022), 109740.

[36] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *Proceedings of the 2018 world wide web conference*. 1277–1286.

[37] Xi Zhang, Akshay Aravamudan, and Georgios C Anagnostopoulos. 2022. Anytime Information Cascade Popularity Prediction via Self-Exciting Processes. In *International Conference on Machine Learning*. PMLR, 26028–26047.

[38] Qingyuan Zhao, Murat A Erdogdu, Hera Y He, Anand Rajaraman, and Jure Leskovec. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1513–1522.

[39] Ting Zhong, Jian Lang, Yifan Zhang, Zhangtao Cheng, Kunpeng Zhang, and Fan Zhou. 2024. Predicting micro-video popularity via multi-modal retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2579–2583.

[40] Fan Zhou, Xin Jing, Xovee Xu, Ting Zhong, Goce Trajcevski, and Jin Wu. 2020. Continual information cascade learning. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 1–6.

[41] 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.

[42] Fan Zhou, Xovee Xu, Kunpeng Zhang, Goce Trajcevski, and Ting Zhong. 2020. Variational information diffusion for probabilistic cascades prediction. In *IEEE INFOCOM 2020-IEEE conference on computer communications*. IEEE, 1618–1627.

A Related Work

Humans naturally leverage multimodal information—including vision, language, and sound—for decision making [1]. Existing multimodal fusion methods often rely on concatenating input modalities using off-the-shelf architectures [12, 19, 26] while several recent works explore MoE for multimodal learning, which offers a natural architecture for multimodal interactions via expert specialization [21, 34]. As for the multimodal information diffusion, researchers have increasingly focused on extending the traditional unimodal popularity prediction task to a multimodal setting. The task of multimodal popularity prediction aims to assess user engagement levels with user-generated images by integrating information from multiple modalities, including image content and textual descriptions. For example, UHAN [36] introduces a user-guided attention framework that combines VGGNet and LSTM to model visual and textual features for predicting image popularity. Similarly, DTCN [31] employs ResNet and LSTM to jointly capture both neighboring and periodic temporal contexts across sequences of images.

However, these works primarily focus on directly predicting the popularity of UGCs based on textual or visual features, while overlooking the temporal diffusion dynamics (i.e., information cascade), which is indeed an essential aspect of information diffusion.

B Datasets

B.1 Data Collection

For the Twitter dataset, we collected all tweets posted between May 1 and June 3, 2021. Original tweets published between May 1 and May 31 were selected as the root nodes of our multimodal cascade graphs, allowing at least three days (from May 31 to June 3) for retweet accumulation. Additionally, cascades containing fewer than 10 nodes were filtered out to ensure sufficient propagation.

For the Twitter-Hashtag dataset, we collected all tweets posted between January 1 and February 28, 2022. Unlike the Twitter dataset, we selected only original tweets whose textual content contained at least one hashtag. An information cascade was then defined based on a shared hashtag between an original tweet and its corresponding retweets. We selected original tweets published between January 1 and January 31, 2022, and included retweets up to February 28, allowing a minimum 28-day propagation window. Similarly, we filtered out multimodal cascade graphs containing fewer than 10 nodes. Note that while our Twitter dataset focuses on the popularity of individual tweets, the Twitter-Hashtag dataset focuses on the popularity of individual hashtags.

For the Weibo dataset, we use the available Weibo data[3], which contains the user, text information, and a 24-hour retweeting chain to form the cascade, but without the image information.

B.2 Dataset Analysis

Table 5 furnishes a comprehensive depiction of each sample within the dataset, representing raw data. Each entry serves as a basis for constructing multimodal cascades, encapsulating the essence of the information embedded in the raw data.

Furthermore, we have tabulated the count of nodes and edges in the global graph at various observation times in Table 6. The observation reveals that the scale of the global graph in the Twitter-hashtag dataset surpasses that of the Twitter and Weibo datasets. This disparity in size could be attributed to the intricate nature of the Twitter-hashtag network. Consequently, the increased values of MSLE and MAPE in the Twitter-hashtag dataset compared to the Twitter and Weibo datasets may stem from this network complexity.

Table 5: Dataset Sample Description

Sample Name	Description
parent_tweet_id	The unique identifier of the parent tweet, indicating the original tweet to which the current tweet is replying.
parent_user_id	The unique identifier of the user who posted the parent tweet.
tweet_id	The unique identifier of the current tweet.
user_id	The unique identifier of the user who posted the current tweet.
time_diff	The time difference between the current tweet and its parent tweet.
text	The textual content of the tweet.
photo	The information or URL link of the photo attached to the tweet.

Table 6: Detailed global graph statistics

Dataset	Twitter	Twitter-Hashtag	Weibo
Nodes(1h/1d/0.5h)	54,116	519,188	74,442
Edges(1h/1d/0.5h)	48,834	769,493	95,154
Nodes(3h/3d/1h)	84,654	792,017	122,710
Edges(3h/3d/1h)	80,486	1,234,015	158,785

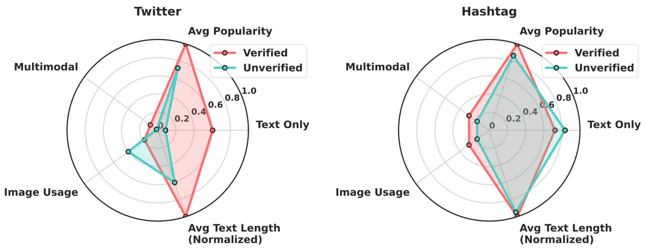


Figure 7: User analysis

What's more, we plot the distribution of popularity and the normalized popularity propagation process from three datasets in Figure 8. Following previous works, we opted for the initial and intermediate phases of cascade diffusion as our observation period, specifically 1 hour and 3 hours for Twitter, 1 day and 3 days for Twitter-Hashtag, 0.5 hours and 1 hour for Weibo.

C Baselines and Metrics

C.1 Baselines

We select eight state-of-the-art baselines of two categories as follows: (1) *UGC-based popularity prediction*: **DTCN** [31] focuses on sequential popularity prediction by integrating visual features and user attributes. **UHAN** [36] constructs a user-guided hierarchical attention network for popularity prediction. (2) *cascade-based methods*: **Feature-based** approaches extract various hand-crafted features and here we use the following features: the size of the observed cascade, the temporal interval between the original node and its initial forwarding, the time at which the last retweet occurs. **DeepHawkes** [2] integrates the Hawkes process and deep learning

[3] http://sklccc.com.cn/

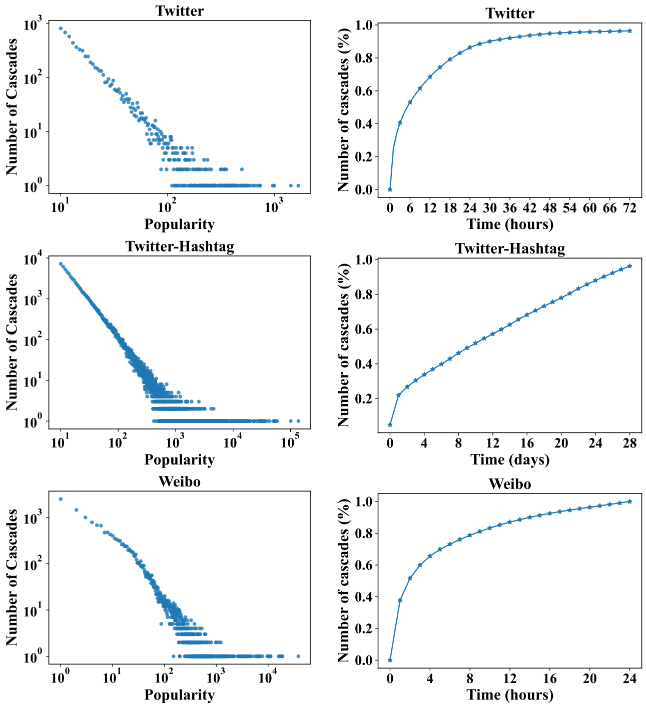


Figure 8: The distribution of popularity and the normalized popularity propagation process.

techniques for the purpose of modeling cascades. **CasCN** [5] samples cascade graph and develops a dynamic graph convolutional network. **CasFlow** [32] mainly considers the effect of both the local graph and global graphs to represent user behavior for predicting popularity. **CTCP** [20] combines all cascades into a diffusion graph and takes the correlation between cascades and the dynamic preferences of users into account. **CasDo** [7] introduces a probabilistic diffusion model to consider the uncertainties in information diffusion. **CasFT** [14] leverages diffusion models to consider the popularity of the future trend.

C.2 Metrics

Mean squared logarithmic error (MSLE) and mean absolute percentage error (MAPE) are defined as follows:

$$MSLE = \frac{1}{M}\sum_{k=1}^{M}(log_2(Y+1) - log_2(\hat{Y}+1))^2 \tag{13}$$

$$MAPE = \frac{1}{M}\sum_{k=1}^{M}\frac{|log_2(Y+2) - log_2(\hat{P}+2)|}{log_2(Y+2)}. \tag{14}$$

D Ablation Study & Hyperparameters

We investigate the influence of hyperparameters, including the hidden dimension after projection. The hidden dimension is varied across values of 16, 32, 64, and 128. The results of the impact of different hidden dimensions are shown in Figure 9. We see that the choice of the hidden dimension has certain influence on the three datasets on both metrics.

Table 7: Combination of two modalities on Twitter (3 hours), Twitter-Hashtag (3 days), and Weibo (1 hour).

Method	Twitter		Twitter-Hashtag		Weibo	
	MSLE	MAPE	MSLE	MAPE	MSLE	MAPE
MMCas-CU	1.3597	0.2554	4.2479	0.4794	2.1543	0.3316
MMCas-CT	1.3651	0.2534	4.2814	0.4814	2.0021	0.3345
MMCas-CV	1.3749	0.2564	4.3111	0.4889	-	-
MMCas-UT	1.6104	0.3143	5.2334	0.5265	2.8795	0.4190
MMCas-UV	1.6241	0.3197	5.4156	0.5314	-	-
MMCas-TV	1.7341	0.3446	5.5178	0.5384	-	-
MMCas	**1.2512**	**0.2105**	**3.4976**	**0.4076**	**1.8015**	**0.3186**

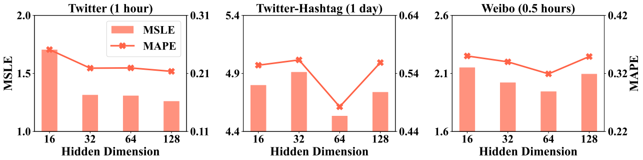


Figure 9: Impact of the hidden dimension.

E Interpretability and Case Study

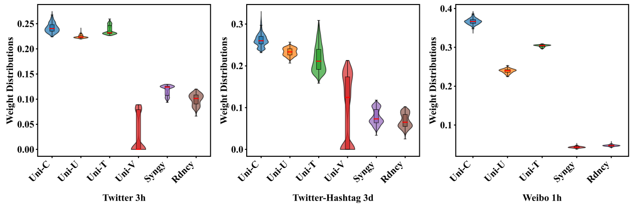
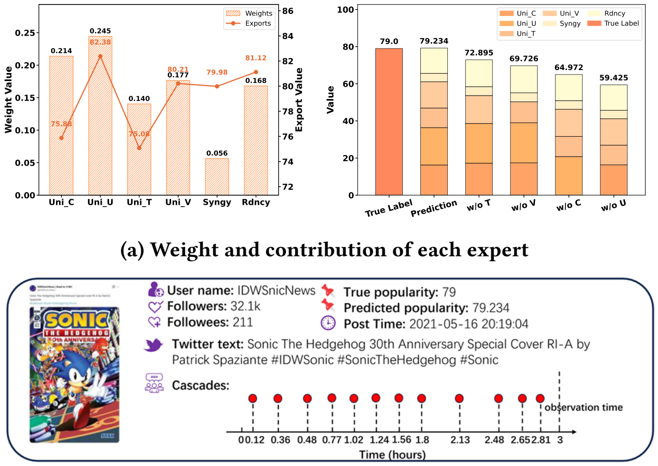


Figure 10: Weight distributions on Twitter (3 hours), Twitter-Hashtag (3 days), and Weibo (1 hour).



(a) Weight and contribution of each expert

User name: IDWSnicNews
Followers: 32.1k
Followees: 211
True popularity: 79
Predicted popularity: 79.234
Post Time: 2021-05-16 20:19:04
Twitter text: Sonic The Hedgehog 30th Anniversary Special Cover RI-A by Patrick Spaziante #IDWSonic #SonicTheHedgehog #Sonic
Cascades:

(b) A multimodal information cascade

Figure 11: Case study