Structural Entropy Guided Incremental Learning for Open-World Multimodal Social Event Detection

Zhiwei Yang[1,2], Haimei Qin[1]*, Xiaoyan Yu[3], Hao Peng[4], Lei Jiang[1], Li Sun[5], Zhiqin Yang[6]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Beijing Institute of Technology, Beijing, China
[4]Beihang University, Beijing, China
[5]North China Electric Power University, Beijing, China
[6] The Chinese University of Hong Kong, Hong Kong, China
yangzhiwei@iie.ac.cn, qinhaimei@iie.ac.cn, xiaoyan.yu@bit.edu.cn, penghao@buaa.edu.cn, jianglei@iie.ac.cn, ccesunli@ncepu.edu.cn, yangzqccc@link.cuhk.edu.hk

Abstract

With the explosive growth of multimodal data streams on social media, the timely detection of emerging social events has become increasingly important. As a result, Multimodal Social Event Detection in open-world settings is receiving growing attention. However, most existing methods face two major limitations: (1) They overlook the dynamic nature of open-world social media data and fail to design dedicated incremental learning frameworks. (2) They ignore the impact of noise in streaming data, leading to performance degradation over long-term detection. To overcome these limitations, we propose SeInEvent (**S**tructural **E**ntropy Guided **In**cremental Learning for Open-World Multimodal Social **Event** Detection). Our innovations are as follows: **First**, considering data dynamics, we design a self-supervised alternating incremental contrastive learning mechanism. Through knowledge distillation, historical event clusters were reviewed and consolidated, and contrastive learning was combined to absorb knowledge of unknown events, ultimately achieving incremental learning without labels. **Second**, addressing the impact of noise, we propose a Pointwise Structural Entropy-based noise filter, which quantifies each sample's informational contribution to the event clustering structure. It enables automatic removal of noisy data and supports robust long-term detection. Extensive experiments on two public datasets demonstrate that SeInEvent achieves superior performance.

Introduction

Social media platforms (e.g., Twitter, Facebook, Instagram) have become major channels for real-time information dissemination and public engagement, continuously producing massive streams of multimodal user-generated content (text, images, videos). These data streams encode valuable signals about real-world social events, such as natural disasters, large gatherings, and cultural celebrations. Accurate and timely detection of such events—characterized by temporal, spatial, and semantic attributes—is essential for understanding social dynamics, analyzing public opinion, and supporting crisis response (Zhang et al. 2024). Early-stage social event detection thus presents a core challenge for intelligent systems in domains including emergency management (Beck et al. 2021), opinion analysis (Pohl, Bouchachia, and Hellwagner 2012), and political decision-making (Marozzo and Bessi 2018). As a result, Multimodal Social Event Detection (MSED) in open-world settings has received growing attention in recent years.

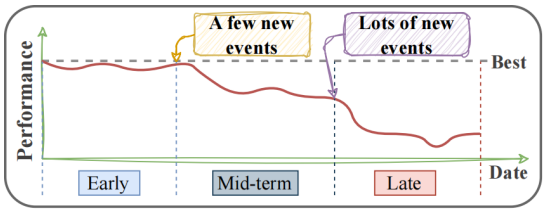

Figure 1: The traditional models during long-term detection: the early model performs well, the mid-term model shows a decline in performance, and the late model performs poorly.

The core task of MSED is to automatically identify and aggregate messages from massive, heterogeneous data sources that refer to the same real-world event, thereby forming semantically coherent event clusters (Zhou et al. 2020). In open-world settings, social media data is continuously generated in a streaming manner, and social events themselves evolve dynamically over time. This necessitates that MSED systems not only capture streaming data effectively but also exhibit lifelong learning capabilities to adapt to both the evolution and emergence of events. Given these requirements, open-world MSED faces two fundamental challenges: **(1) The high dynamics of data in the open world.** Social events in open-world environments are inherently dynamic, with newly emerging events often displaying previously unseen characteristics. Without incremental learning capabilities, models rapidly become outdated (as illustrated in Figure 1). Most existing SED methods (Cao et al. 2021; Peng et al. 2021) rely on supervised learning with static datasets, which is suitable for offline, one-time

*Corresponding author

training but poorly suited to streaming scenarios. A recent study (Cao et al. 2024) employs clustering to handle streaming data but fails to leverage insights from historical events. By contrast, incremental learning—which continuously updates the model—offers a promising direction. Traditional incremental methods (Qian et al. 2025; Jin et al. 2021) mitigate catastrophic forgetting by maintaining a buffer of historical samples. However, these approaches remain inadequate in the context of open-world MSED, where the primary focus lies not in remembering past samples but in effectively detecting novel events. The core problem of incremental learning in open-world MSED is that, after the model has been exposed to many events, it tends to confuse the representations of different events, leading to blurred decision boundaries and degraded performance. **(2) The challenge of noise interference in the open world.** Open-world social media streams are inundated with noisy content. Incorporating such noise into the incremental learning process contaminates the model's representation space, blurs cluster boundaries, reduces cluster purity, and progressively impairs the accuracy of novel event detection. Manual noise filtering is impractical and incompatible with the automation and real-time requirements of streaming data processing in open-world scenarios. Yet, current research in both incremental learning and the SED domain has largely overlooked the pervasive presence of noise in streaming data. Consequently, devising mechanisms to automatically filter noise from streaming data remains a critical challenge for open-world MSED.

Facing these challenges, we propose SeInEvent, a self-supervised framework designed to support continual MSED in open-world environments. **To address Challenge 1**, we introduce a fully self-supervised alternating incremental contrastive learning framework that interleaves knowledge distillation-based retrospection with new knowledge acquisition. The retrospection phase distills knowledge from the current model into the updated model, preserving historical event cluster representations. In parallel, the acquisition phase enhances the model's understanding of emerging events via self-supervised contrastive learning on newly arrived data. This framework operates entirely without manual annotations throughout both pretraining and all incremental stages. **To address Challenge 2**, we propose a novel Pointwise Structural Entropy (PSE)-based noise filtering algorithm. By quantifying each sample's information contribution within the clustering graph, PSE dynamically filters out outliers and low-information noise, retaining only high-quality core samples for incremental learning. Through these mechanisms, SeInEvent enables robust MSED with long-term cluster stability and effective knowledge retention in open-world scenarios.

Our main contributions are as follows: 1) We propose SeInEvent, an incremental learning framework designed for open-world, multimodal, and streaming social event detection. To the best of our knowledge, the open-world MSED has not been studied. 2) We propose a novel self-supervised alternating incremental contrastive learning strategy. It consolidates historical event clusters via knowledge distillation and learns emerging events through contrastive learning on new data, achieving stable knowledge retention and continual adaptation without manual annotations. 3) We first propose a PSE-based noise filtering mechanism, which dynamically quantifies the informational value of each sample within the clustering structure, automatically identifies and removes noisy instances, and significantly enhances the robustness of incremental learning as well as the purity of event clusters.

Preliminary

Task Definition

Given a set of social messages $M = \{m_1, \ldots, m_n\}$, where each message $m_n = \{T_n, I_n\}$ is a text-image pair, the goal of MSED is to partition M into a set of disjoint clusters $P = \{p_1, \ldots, p_j\}$, where each cluster p_j represents a distinct social event. These clusters satisfy $p_i \cap p_j = \emptyset$ for $i \neq j$, and jointly cover all messages: $p_1 \cup \ldots \cup p_j = M$.

Encoding Tree and Structural Entropy

Structural Entropy (SE) (Li, Li, and Pan 2015; Li and Pan 2016) is a graph structure information measurement indicator based on an encoding tree. SE has a good effect in measuring clustering quality (Sun et al. 2024).

Given a graph $\mathcal{G} = (V, E)$, the encoding tree $\mathcal{T}$ includes all nodes V as leaf nodes. Each node α in $\mathcal{T}$ corresponds to a partitioning of message nodes, with the set $\mathcal{T}_\alpha = v_a^1, \ldots, v_a^j$, representing the successor nodes of α. The root node λ of $\mathcal{T}$ has the set $\mathcal{T}_\lambda = V$, indicating no partitioning. For each node α in $\mathcal{T}$ (excluding λ), the height $h(\alpha)$ is one less than that of its parent node. The root node λ has a height of 0. The height of $\mathcal{T}$ is the maximum height among all nodes in $\mathcal{T}$.

SE is calculated based on the encoding tree, and any encoding tree $\mathcal{T}$ corresponds to its SE. The 2-dimensional SE (2D SE) represents the stability of the node partition in graph G and also represents the quality of clustering (The smaller 2D SE is, the better.). 2D SE is calculated as follows:

$$H^{(2)}(\mathcal{G}) = -\sum_{j=1}^{m} \frac{V_j}{w} \sum_{i=1}^{n_j} \frac{d_i^j}{V_j} \log_2 \frac{d_i^j}{V_j} - \sum_{j=1}^{m} \frac{P_{cut_j}}{w} \log_2 \frac{V_j}{w}, \tag{1}$$

where n_j is the number of nodes in partition e_j, d_i^j is the weighted degree of the i-th node in e_j, V_j is the sum of the weighted degrees of all nodes in partition e_j, P_{cut_j} is the sum of the weights of the cut edges in e_j, and w denotes the sum of the weighted degrees of all nodes.

Pointwise Structural Entropy

PSE usually indicates the uncertainty or complexity of the information surrounding a particular node. For a node in partition $\mathcal{T}_{\alpha_j}$, whose PSE is calculated as:

$$\mathcal{H}^v(\mathcal{G}) = -\frac{d_i^j}{V_j} \log_2 \frac{d_i^j}{V_j}. \tag{2}$$
$$v \in \mathcal{T}_{\alpha_j}$$

The larger PSE indicates that the node may not belong to the current community or even the node may belong to the noisy data.

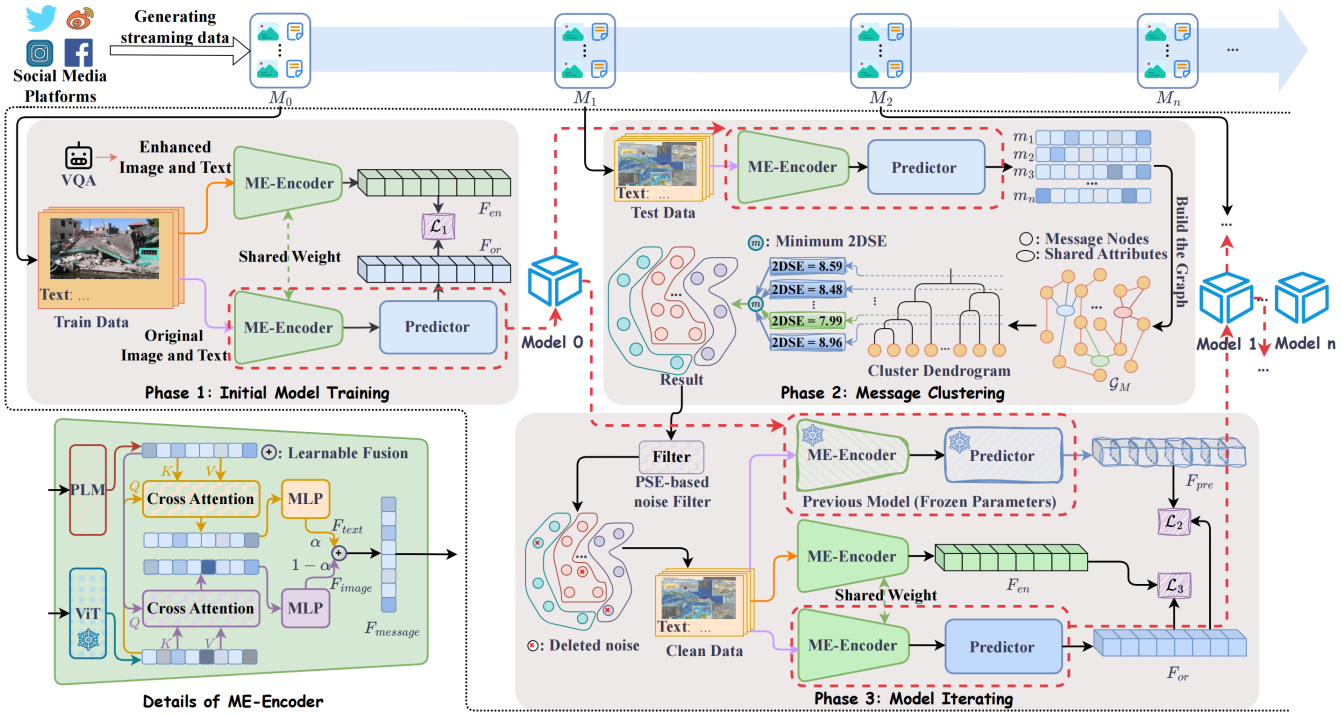

Figure 2: Detailed design of our proposed SeInEvent. In Phase 1, self-supervised contrastive learning is employed to pretrain the model and obtain robust representations. In Phase 2, a clustering algorithm based on SE is applied for MSED. In Phase 3, noise filtering is first conducted on incoming data, followed by alternating incremental contrastive learning to update the model.

Methodology

We describe SeInEvent in detail, as shown in Figure 2.

Initial Model Training

Positive Sample Construction. To ensure SeInEvent operates without reliance on supervised labels, we adopt a label-free contrastive learning strategy without negative samples, a proven approach in prior work (Grill et al. 2020; Chen and He 2021). Positive samples are generated via a Multimodal Large Language Model (MLLM) using Visual Question Answering (VQA) on original messages to reconstruct semantically aligned pairs. To enrich the construction of positive samples, we focus on three key attributes of a multimodal message: **event type** (e.g., sports, disaster, conflict), **event theme** (e.g., post-earthquake recovery, coastal evacuation), and **image caption**, which integrates visual context. Subsequently, the answers from MLLMs are concatenated with the original text to form enhanced text. *See the Appendix for the prompts.* We employ InstructBLIP (Li et al. 2023) to enhance textual content from both modalities:

$$T_{en} = \text{BLIP}(T_{or}, I_{or}), \tag{3}$$

where T_{or} and I_{or} denote the original text and image, and T_{en} is the enhanced text.

Multimodal Event Encoder. To transform multimodal messages into unified vector representations, we employ a Multimodal Event Encoder (ME-Encoder), illustrated in Figure 2. First, text embeddings are extracted using SBERT (Reimers and Gurevych 2019), and image features are obtained via a frozen pre-trained Vision Transformer (ViT) (Dosovitskiy 2020). Then, a bidirectional cross-attention mechanism facilitates interaction between these text and image features. Finally, the attended features are adaptively fused into the final message embedding $F_{message}$ using a learnable weight α:

$$F_{message} = \alpha \cdot F_{text} + (1 - \alpha) \cdot F_{image}, \tag{4}$$

where $0 < \alpha < 1$ is learnable. This process enables the proposed ME-Encoder to produce a compact representation that captures the multimodal information within each social media message, suitable for subsequent event detection tasks.

Training Objective. To eliminate the reliance on sample labels, we adopt a self-supervised contrastive learning approach without using negative samples. Specifically, both the enhanced and original data are passed through two weight-sharing ME-Encoders. In order to facilitate the stabilization of the training model, a two-layer multi-layer perceptron predictor was added to the branch that processes the raw data. Contrastive learning is then employed to maximize the similarity between the outputs of the ME-Encoders:

$$\mathcal{L}_1 = -\frac{F_{or}}{\|F_{or}\|_2} \cdot \frac{F_{en}}{\|F_{en}\|_2}. \tag{5}$$

where F_{or} denotes the ME-Encoder output for the original data after passing through the predictor, and F_{en} denotes the output for the enhanced data.

Algorithm 1: Clustering Based on SE

Input: Message graph $\mathcal{G}_M$ with n nodes.
Output: Clustering result P.

1: $\mathcal{T}_h \leftarrow$ Run hierarchical clustering and obtain a cluster tree.
2: Set $SE = \emptyset$, set $i = 1$.
3: **while** $i < n$ **do**
4: $P_i \leftarrow$ Obtain the clustering results P_i under the clusters for i according to $\mathcal{T}_h$.
5: $se_i \leftarrow$ Calculate the 2DSE of P_i according to Eq. (1).
6: $SE \overset{\text{add to}}{\leftarrow} se_i$
7: $i = i + 1$
8: **end while**
9: i_{bset} = arg max$(SE) + 1$
10: $P = P_{i_{bset}} \leftarrow$ Optimal clustering results.
11: **return** P

Algorithm 2: PSE-based noise filtering

Input: Message graph $\mathcal{G}_N$ with noise. Detection results P. Hyperparameter μ $(0 < \mu < 1)$.
Output:$\mathcal{G}_C$ without noise.

1: **for** each p_i in P **do**
2: Set $PSE = \emptyset$. Set $\mathcal{G}_C = \mathcal{G}_N$.
3: **for** each m_i in p_i **do**
4: $ne_{m_i} \leftarrow$ Calculate PSE based on Eq. (2).
5: $PSE \overset{\text{add to}}{\leftarrow} ne_{m_i}$.
6: **end for**
7: $PSE_{aver} = \{x \mid x \in PSE \text{ and } x > \text{mean}(PSE)\}$
8: $a = \text{average}(PSE_{aver}) \cdot (1 + \mu)$.
9: **for** each ne_{m_i} in PSE **do**
10: **if** $ne_{m_i} > a$ **then**
11: $\mathcal{G}_C$ = $\mathcal{G}_C$ - m_i.
12: **end if**
13: **end for**
14: **end for**
15: **return** $\mathcal{G}_C$

Message Clustering

This process is inspired by HISEvent (Cao et al. 2024), which first constructs a graph and then applies graph clustering. In contrast, while HISEvent adopts a greedy iterative search for the optimal encoding tree, we integrate hierarchical clustering to eliminate such costly iterations, resulting in improved efficiency.

Message Graph Building. We model it as a graph ($\mathcal{G}_M$) with two types of edges: feature similarity-based edges (E_s) and shared attribute-based edges (E_a). For E_s, each node is connected to the node with the most similar PLM output. Edge weights are determined by the cosine similarity between the connected nodes. For E_a, nodes are linked based on shared attributes, such as posts from the same user, posts with a common hashtag (#), or posts mentioning the same account (@).

Clustering Based on SE. To cluster messages without predefining the number of events, we employ the Clustering Based on SE algorithm, as detailed in Algorithm 1. The algorithm takes the constructed message graph $\mathcal{G}_M$ as input. First, it applies a hierarchical clustering algorithm to generate a hierarchical clustering tree ($\mathcal{T}_h$) (Line 1). Each layer of $\mathcal{T}_h$ corresponds to a clustering result, for which we compute the 2DSE (Lines 3–8). The clustering result with the smallest 2DSE is selected as the final output (Lines 9–11), as a lower 2DSE indicates better clustering quality.

Model Iterating

After MSED, SeInevent will perform model iteration. Directly using all of them for incremental learning may blur cluster boundaries and degrade cluster purity. To address this, we first apply PSE-based noise filtering, followed by alternating incremental contrastive learning.

PSE-based noise filtering. To filter out noise and low-quality data, we propose a novel *PSE-based noise filtering* algorithm, as outlined in Algorithm 2. For each cluster in $\mathcal{G}_N$, the process proceeds as follows. First, we compute the node entropy for each node in the cluster and store it in the set PSE (Lines 3-6). Next, we determine the threshold a using PSE_{aver} (Lines 7-8), where PSE_{aver} consists of all values in PSE that exceed the average entropy. Finally, nodes with entropy above the threshold are identified as noise and removed to refine the dataset (Lines 9-13).

Alternating Incremental Contrastive Learning. To keep SeInEvent up-to-date, we introduce an alternating incremental contrastive learning, as illustrated in Phase 3 of Figure 2. New messages are first filtered to remove noise, yielding clean inputs for learning. Unlike initial training, incremental learning must preserve prior knowledge. To this end, we employ a distillation module (see *Previous Model* in Phase 2 of Figure 2) that transfers knowledge from the frozen ME-Encoder and Predictor of the previous model. Meanwhile, the two shared-weight ME-Encoders are reused from Phase 1. During training, each batch of clean data is split: 50% is used for distillation with the previous model, and 50% for acquiring new knowledge through the current ME-Encoder. Extensive experiments show that fixing the data ratio and adjusting the balance between the two processes via their loss weights leads to more stable model training.

Iterating Objective. The adaptive incremental learning phase incorporates two loss functions: the distillation loss ($\mathcal{L}_2$), which preserves previously learned knowledge, and the contrastive loss ($\mathcal{L}_3$), which facilitates the acquisition of new knowledge. To enhance the model's learning process, we unify these two losses as follows:

$$\mathcal{L}_2 = -\frac{F_{or}}{\|F_{or}\|_2} \cdot \frac{F_{pre}}{\|F_{pre}\|_2}, \mathcal{L}_3 = -\frac{F_{or}}{\|F_{or}\|_2} \cdot \frac{F_{en}}{\|F_{en}\|_2}, \quad (6)$$

where F_{pre} represents the output of the previous model. Ultimately, the total loss ($\mathcal{L}_{In}$) is the sum of the two losses mentioned above:

$$\mathcal{L}_{In} = \beta \cdot \mathcal{L}_2 + (1 - \beta) \cdot \mathcal{L}_3, \quad (7)$$

where β is a hyperparameter used to control the weights of the two types of losses, and $0 < \beta < 1$.

Metrics	BERT	SBERT	HISEvent	MMBT	SCBD	CLIP	OWSEC	BLIP2	LLaVA	MFEK	ODII	SeInEvent	Improv. (%)
ARI	.03	.59	.72	.58	.73	.79	.74	.60	.65	.75	.72	**.80**	↑1
NMI	.07	.67	.76	.57	.78	.76	.80	.68	.69	.79	.75	**.81**	↑1
AMI	.07	.66	.75	.57	.78	.75	.80	.68	.69	.79	.76	**.81**	↑1

Table 1: (RQ1) MSED in the closed world. The best results are bolded, and the second-best results are underlined.

Methods	M_1			M_2			M_3			M_4			M_5			M_6			M_7		
	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI
BERT	.15	.15	.15	.22	.29	.29	.08	.18	.17	.15	.24	.24	.21	.21	.21	.30	.39	.39	.09	.13	.12
SBERT	.30	.44	.44	.44	.52	.52	.21	.38	.37	.45	.62	.62	.53	.65	.65	.67	.71	.71	.30	.45	.45
HISEvent	.18	.40	.29	.19	.40	.28	.10	.39	.23	.55	.72	.71	.42	.61	.59	.22	.41	.29	.16	.44	.43
MMBT	.12	.10	.10	.11	.10	.10	.12	.13	.12	.09	.08	.08	.10	.11	.12	.15	.12	.13	.05	.08	.08
SCBD	.19	.24	.24	.28	.30	.30	.16	.21	.21	.19	.26	.26	.10	.15	.15	.25	.24	.24	.05	.08	.09
CLIP	.64	.67	.67	.70	.70	.70	.30	.49	.49	.54	.69	.69	**.86**	**.85**	**.84**	.84	.85	.85	.09	.20	.19
OWSEC	.21	.23	.23	.29	.32	.32	.15	.19	.19	.19	.27	.26	.15	.17	.16	.27	.25	.25	.02	.09	.09
BLIP2	.42	.44	.44	.39	.39	.38	.26	.13	.12	.11	.19	.19	.21	.30	.26	.37	.41	.41	.15	.16	.18
LLaVA	.45	.46	.46	.40	.42	.42	.30	.48	.48	.30	.31	.31	.44	.43	.43	.45	.48	.48	.20	.23	.23
MFEK	.34	.36	.36	.29	.30	.30	.30	.30	.30	.18	.25	.25	.16	.15	.17	.29	.26	.25	.07	.18	.18
ODII	.33	.32	.32	.28	.30	.29	.25	.27	.27	.14	.13	.13	.16	.15	.15	.28	.28	.28	.03	.10	.11
SeInEvent	**.95**	**.93**	**.93**	**.99**	**.97**	**.97**	**.97**	**.93**	**.93**	**.91**	**.88**	**.88**	.82	.81	.81	**.96**	**.94**	**.94**	**.32**	**.49**	**.50**
Improv.(%)	↑48	↑39	↑39	↑41	↑39	↑39	↑223	↑90	↑90	↑69	↑28	↑28	↓5	↓5	↓4	↑14	↑11	↑11	↑60	↑113	↑117

Table 2: (RQ1) MSED in the open world. The best results are bolded, and the second-best results are underlined.

Experiments

We conduct comprehensive experiments to evaluate SeInEvent. Specifically, we aim to answer the following research questions: **RQ1:** Compared to existing methods, how does SeInEvent perform on MSED tasks in offline and incremental scenarios? **RQ2:** How does each proposed module contribute to the overall performance? **RQ3:** How robust is SeInEvent when facing different hyperparameters or MLLMs choices? **RQ4:** In latent space, does the proposed incremental strategy learns a clear clustering boundary?

Experimental Setup

Datasets. We use two public datasets: CrisisMMD (Alam, Ofli, and Imran 2018) and NED (Lin, Xie, and Li 2024). CrisisMMD consists of 7 events, making it more suitable for MSED in a closed environment due to its limited event diversity. Thus, we apply stratified sampling to divide it into a training set, validation set, and test set in a 7:1:2 ratio. NED comprises 40 events spanning ten years, making it ideal for simulating incremental learning scenarios in an open-world setting. We define the first three years as the initial message block M_0, while each subsequent year from the fourth onward is treated as a separate message block $M_1, \ldots, M_7$.

Baselines. We compare SeInEvent against two categories of baselines. **Unimodal baselines**: **BERT** (Devlin et al. 2019): A pre-trained language model fine-tuned on our dataset. **SBERT**: A sentence-transformer model with enhanced semantic representation, also fine-tuned on our data. We perform SED by applying K-means clustering to their output features. **HISEvent**: An unsupervised method based on structural entropy minimization, representing the current state-of-the-art in unimodal event detection. **Multimodal baselines**: **MMBT** (Kiela et al. 2019): Uses a Transformer to fuse textual and visual features for classification. **SCBD** (Abavisani et al. 2020): Fuses multimodal features via a cross-attention mechanism. **CLIP** (Radford et al. 2021): Trained with contrastive learning on large-scale image-text pairs; we use its fused image-text embeddings. **OWSEC** (Qian et al. 2023): An open-world multimodal event classification model using masked learning. **BLIP2** and **LLaVA** (Liu et al. 2024): Two pre-trained MLLMs; we use pooled representations from their final hidden layers. **MFEK** (Lin, Xie, and Li 2024): A closed-world method leveraging external knowledge. **ODII** (Yu, Hu, and Wang 2025): An open-world model that performs disaster identification via multi-task learning. To ensure fairness, we train MMBT, SCBD, OWSEC, MFEK and ODII under supervised settings and use their feature embeddings for K-means clustering during evaluation.

Implementation Details. The proposed framework is implemented in PyTorch, and all experiments are conducted on a Linux server equipped with 8 RTX 3090 GPUs. For SeInEvent, we set the batch size to 64 and train for 50 epochs during initial training. In the incremental phase, the batch size is increased to 128, with 30 training epochs, using $\beta = 0.5$ and $\mu = 0.5$. The model is first trained on M_0 and evaluated on M_1, then updated with M_1 and tested on M_2, and so on. For baselines, the training set for evaluating M_i includes all data from M_0 to M_{i-1}. For K-means, the number of clusters is set to the true number of events.

RQ1: Overall Performance

We evaluate SeInEvent in both closed-world and open-world MSED. Results for SeInEvent and baselines are reported in Table 1 and Table 2. All results are averages of five runs.

Closed World. In the closed world, SeInEvent delivers the best overall performance. Among unimodal baselines, BERT performs the worst, SBERT offers moderate gains, and HISEvent leads unimodal models but still trails behind multimodal methods, underscoring the importance of multimodal fusion in SED. Despite extensive pretraining, MLLMs perform poorly, suggesting limited effectiveness in event-level modeling. CLIP surpasses unimodal baselines, likely due to its contrastive cross-modal training. SCBD and MFEK perform well, proving effective in closed settings, while OWSEC, with its masked fusion strategy, stands out as the strongest supervised baseline. Unlike these methods,

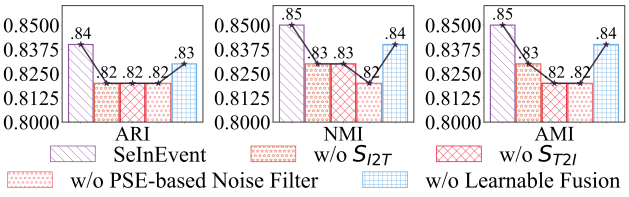

Figure 3: (RQ2) Results of ablation experiments. Values are the average of all message blocks.

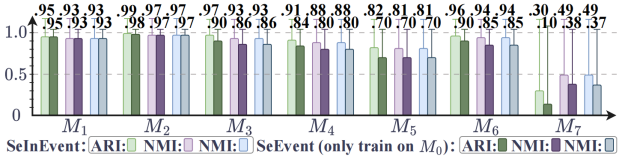

Figure 4: (RQ2) Results of experiments about incremental learning analysis. SeEvent is a degenerate model of SeInEvent, i.e., no incremental learning is performed and only the initial message M_0 is used for training.

SeInEvent is self-supervised and does not require labels or predefined event counts, making it more scalable and suitable for real-world deployment.

Open World. In the open world, SeInEvent again leads overall, ranking first in most message blocks and narrowly trailing CLIP on M_5. Unimodal models follow similar patterns as in the closed setting. Supervised methods like MMBT, SCBD, OWSEC, and MFEK see notable performance declines, revealing their dependence on training distributions and limited generalization to novel events. While BLIP2 and LLaVA show slight improvements on M_7, their performance remains inconsistent. CLIP performs well on select blocks, benefiting from large-scale contrastive training. SeInEvent, without retraining or storing historical data, maintains robust performance on evolving data streams, thanks to its alternating incremental contrastive learning and PSE-based noise filtering.

RQ2: Ablation Study

Ablation of Components. SeInEvent comprises four key components: image-to-text attention (S_{I2T}), text-to-image attention (S_{T2I}), a PSE-based noise Filter, and Learnable Fusion within the ME-Encoder. To evaluate their contributions, we conduct ablation studies by removing each component separately, denoted as 'w/o X', while keeping other settings unchanged. For 'w/o Learnable Fusion', we replace the fusion module with simple averaging. As shown in Figure 3, all components contribute to performance. In particular, the PSE-based noise Filter proves essential in mitigating noise during incremental learning, leading to more efficient training. The bidirectional attention mechanism facilitates richer cross-modal feature interaction, capturing both global context and fine details. Learnable fusion further improves feature integration, highlighting its importance in enhancing representation quality.

models	M_1	M_2	M_3	M_4	M_5	M_6	M_7
Trained on M0	0.9367	-	-	-	-	-	-
Incremental on M1	0.9367	0.9767	-	-	-	-	-
Incremental on M2	0.9367	0.9767	0.9433	-	-	-	-
Incremental on M3	0.9367	0.8894	0.9131	0.8900	-	-	-
Incremental on M4	0.9391	0.9770	0.9287	0.8622	0.8133	-	-
Incremental on M5	0.9391	0.8644	0.9360	0.8842	0.8123	0.9467	-
Incremental on M6	0.9391	0.9767	0.9433	0.8900	0.8123	0.9495	0.4267

Table 3: (RQ2) Results of robustness experiments. Values are the average of all metrics.

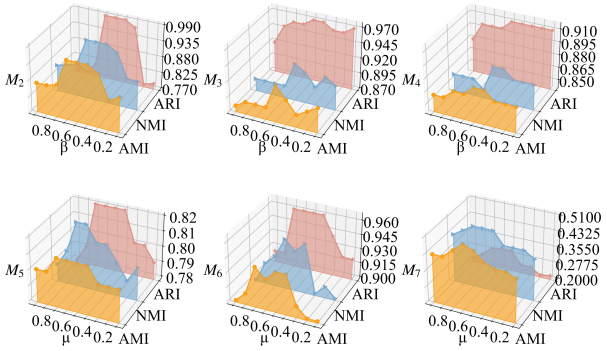

Figure 5: (RQ3) Results of hyperparameter sensitivity.

Analysis of Robustness. To evaluate the robustness of SeInEvent, particularly its ability to retain prior knowledge across successive iterations, we conduct experiments using the current model on earlier message blocks, as shown in Table 3. The results demonstrate that SeInEvent effectively preserves prior knowledge despite undergoing multiple rounds of incremental learning. In most cases, the model maintains performance comparable to previous iterations, with only minor declines observed in certain instances. Notably, SeInEvent appears to reinforce its understanding of earlier knowledge through the incremental learning process. For example, in experiments involving M_1 and M_6, the model from the most recent iteration not only retained its prior knowledge but also exhibited improved performance. This suggests that SeInEvent not only preserves existing knowledge but also promotes continual improvement in comprehension, underscoring its robustness in dynamic environments.

Analysis of Incremental Learning. To assess the effectiveness of SeInEvent in long-term detection, we compare SeInEvent with SeEvent, which is trained only on the initial message block. As shown in Figure 4, both methods perform similarly at the start. However, as new message blocks arrive, SeEvent's performance deteriorates due to its inability to adapt to evolving data distributions, resulting in a widening performance gap. In contrast, SeInEvent maintains stable accuracy over time, demonstrating the importance of incremental learning for long-term robustness in dynamic environments.

RQ3: Operational Robustness

Hyperparameter Sensitivity. The hyperparameters β and μ are critical to SeInEvent's incremental learning and noise

MLLMs (Parameters)	M_1			M_2			M_3			M_4			M_5			M_6			M_7		
	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI	ARI	NMI	AMI
InstructBLIP (7B)	**.95**	.93	.93	**.99**	**.97**	**.97**	**.97**	**.93**	.93	**.91**	**.88**	**.88**	**.82**	**.81**	**.81**	**.96**	**.94**	**.94**	.32	.49	**.50**
LLaVA-V1.6 (7B)	.90	.86	.86	.85	.82	.83	.93	.90	.90	.89	.85	.87	.80	.80	.80	.94	.93	.92	.33	.44	.45
DeepSeek-VL (7B)	.92	.88	.88	.93	.89	.90	.95	.92	.92	.90	.87	**.88**	**.82**	.80	.80	.95	**.94**	.92	.31	.45	.45
Qwen2-VL (7B)	**.95**	**.94**	**.94**	.96	.93	.93	.95	**.93**	**.94**	**.91**	**.88**	.87	**.82**	.80	.80	**.96**	.93	**.94**	**.34**	**.50**	**.50**
MiniCPM-V2 (3B)	.91	.87	.87	.92	.89	.89	.92	.90	.90	.88	.85	.86	.79	.78	.78	.92	.90	.90	.32	.40	.41

Table 4: (RQ3) Results of using different MLLMs. Other hyperparameters remain consistent with the main experiment.

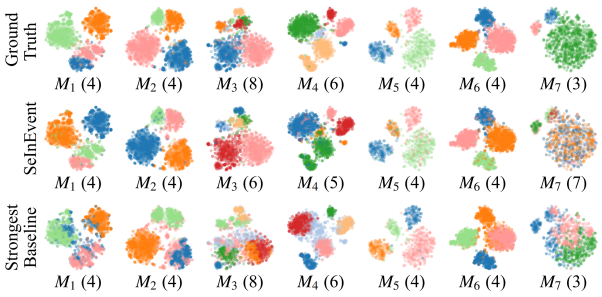

Figure 6: (RQ4) Comparison of SeInEvent clustering results with actual results and the strongest baseline. The number of clusters is indicated in parentheses.

filtering. Specifically, β controls the trade-off between retaining prior knowledge and incorporating new information during model updates, while μ determines the strength of noise suppression. We investigate their impact through a series of experiments, as shown in Figure . Results show that both parameters significantly influence performance. For β, accuracy initially improves with increasing values but declines beyond a certain point, indicating an optimal range for balancing knowledge retention and adaptation. Empirically, we recommend setting β between 0.3 and 0.7. A similar rise-then-fall trend is observed for μ, with optimal values varying across datasets. This aligns with intuition: overly aggressive filtering may discard useful information, while weak filtering fails to suppress noise. Hence, dataset-specific tuning of μ is essential for optimal performance.

Analysis of MLLMs. We use several public MLLMs with identical prompts to generate positive samples and evaluate their effect on final performance, as shown in Table 4. Overall, InstructBLIP (7B) and Qwen2-VL (7B) achieve the best results across most datasets, proving their strength in producing high-quality positives. InstructBLIP (7B) excels on M_1–M_6, while Qwen2-VL (7B) slightly outperforms it on M_7. Other models, such as DeepSeek-VL (7B) and LLaVA-V1.6 (7B), are competitive but slightly behind. These results show that the choice of MLLM affects positive-sample quality and overall performance.

RQ4: Visualization

Figure 6 illustrates SeInEvent's feature space representation and clustering performance. Except for M_7, SeInEvent's clusters align closely with actual outcomes, showcasing its robustness. For M_7, severe class imbalance in event distribution poses challenges, complicating detection.

Related Work

Early Social Event Detection (SED) methods were primarily text-based. Rule-based approaches (Fung et al. 2005; Fedoryszak et al. 2019; Singh and Kumari 2021; Hu et al. 2022) relied on predefined patterns but lacked adaptability and achieved low accuracy. With the emergence of pre-trained language models, text-based SED (Chen et al. 2018; Sihem Sahnoun and Yahia 2020; Li et al. 2024; Yu et al. 2025a) improved via enhanced semantic representations. Graph-based methods (Peng et al. 2019, 2021; Ren et al. 2021; Cao et al. 2021; Ren et al. 2022, 2024; Yang et al. 2024; Yu et al. 2024, 2025b; Zhang et al. 2025) further advanced detection by modeling message relations with GNNs, capturing complex event structures. However, all unimodal methods are limited by their inability to fully exploit multimodal social media data.

MSED has received less attention. KGE-MMSLDA (Xue et al. 2019) combines multimodal topic modeling with external knowledge. SCBD (Abavisani et al. 2020) and OWSEC (Qian et al. 2023) adopt cross-modal fusion techniques to improve classification performance, with OWSEC particularly targeting unseen events in open-world settings. MFEK (Lin, Xie, and Li 2024) further enhances fusion by jointly incorporating external knowledge, text, and images. However, these MSED methods generally treat event detection as a closed-set classification task, limiting their ability to detect novel or evolving events. Moreover, none of them support incremental learning, making them vulnerable to performance degradation over time. In contrast, our proposed SeInEvent enables open-world MSED without labeled data. It supports the detection of unseen events and incorporates adaptive incremental learning, ensuring long-term robustness in evolving social media streams.y

Conclusion

In this paper, we propose a novel Structural Entropy Guided Incremental Learning for Open-World MSED. Specifically, we propose an alternating incremental contrastive learning strategy that interleaves retrospection via knowledge distillation with the acquisition of new event knowledge. This approach not only preserves stable representations of historical event clusters but also enhances the model's ability to capture emerging events. Additionally, we introduce a novel PSE-based noise filtering algorithm, which automatically removes noisy samples from incoming data and retains high-quality core samples for continual updates. Experiments on real-world open datasets demonstrate the effectiveness and practicality of SeInEvent. In future work, we aim to extend MSED to trend prediction, offering deeper insight into event dynamics and enhancing real-world applicability.

Acknowledgments

This research is supported by the National Key R&D Program of China through grant 2023YFC3303800, NSFC through grants 62322202, 62441612 ,62432006 and 62202164, Beijing Natural Science Foundation through grant L253021, Local Science and Technology Development Fund of Hebei Province Guided by the Central Government of China through grants 246Z0102G and 254Z9902G, Major Science and Technology Special Projects of Yunnan Province through grants 202502AD080012 and 202502AD080006, and the Fundamental Research Funds for the Central Universities, and Procurement Project through grant E5V01511D3.

References

Abavisani, M.; Wu, L.; Hu, S.; Tetreault, J.; and Jaimes, A. 2020. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14679–14689.

Alam, F.; Ofli, F.; and Imran, M. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*.

Beck, T.; Lee, J.-U.; Viehmann, C.; Maurer, M.; Quiring, O.; and Gurevych, I. 2021. Investigating label suggestions for opinion mining in German Covid-19 social media. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1–13.

Cao, Y.; Peng, H.; Wu, J.; Dou, Y.; Li, J.; and Yu, P. S. 2021. Knowledge-preserving incremental social event detection via heterogeneous gnns. In *Proceedings of the Web Conference 2021*, 3383–3395.

Cao, Y.; Peng, H.; Yu, Z.; and Philip, S. Y. 2024. Hierarchical and incremental structural entropy minimization for unsupervised social event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8255–8264.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Chen, X.; Zhou, X.; Sellis, T.; and Li, X. 2018. Social event detection with retweeting behavior correlation. *Expert Systems with Applications*, 114: 516–523.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fedoryszak, M.; Frederick, B.; Rajaram, V.; and Zhong, C. 2019. Real-time event detection on social data streams. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2774–2782.

Fung, G. P. C.; Yu, J. X.; Yu, P. S.; and Lu, H. 2005. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*, 181–192.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Hu, X.; Ma, W.; Chen, C.; Wen, S.; Zhang, J.; Xiang, Y.; and Fei, G. 2022. Event detection in online social network: Methodologies, state-of-art, and evolution. *Computer Science Review*, 46: 100500.

Jin, X.; Lin, B. Y.; Rostami, M.; and Ren, X. 2021. Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning. *arXiv preprint arXiv:2104.08808*.

Kiela, D.; Bhooshan, S.; Firooz, H.; Perez, E.; and Testuggine, D. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Li, A.; Li, J.; and Pan, Y. 2015. Discovering natural communities in networks. *Physica A: Statistical Mechanics and its Applications*, 436: 878–896.

Li, A.; and Pan, Y. 2016. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory*, 62(6): 3290–3339.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, P.; Yu, X.; Peng, H.; Xian, Y.; Wang, L.; Sun, L.; Zhang, J.; and Yu, P. S. 2024. Relational Prompt-based Pre-trained Language Models for Social Event Detection. *ACM Transactions on Information Systems*.

Lin, Z.; Xie, J.; and Li, Q. 2024. Multi-modal news event detection with external knowledge. *Information Processing & Management*, 61(3): 103697.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Marozzo, F.; and Bessi, A. 2018. Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining*, 8: 1–13.

Peng, H.; Li, J.; Gong, Q.; Song, Y.; Ning, Y.; Lai, K.; and Yu, P. S. 2019. Fine-grained event categorization with heterogeneous graph convolutional networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3238–3245.

Peng, H.; Li, J.; Song, Y.; Yang, R.; Ranjan, R.; Yu, P. S.; and He, L. 2021. Streaming social event detection and evolution discovery in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5): 1–33.

Pohl, D.; Bouchachia, A.; and Hellwagner, H. 2012. Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st international conference on world wide web*, 683–686.

Qian, S.; Chen, H.; Xue, D.; Fang, Q.; and Xu, C. 2023. Open-world social event classification. In *Proceedings of the ACM Web Conference 2023*, 1562–1571.

Qian, S.; Zhang, S.; Xue, D.; Zhang, H.; and Xu, C. 2025. Learning Temporal Event Knowledge for Continual Social Event Classification. *IEEE Transactions on Knowledge and Data Engineering*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Ren, J.; Jiang, L.; Peng, H.; Cao, Y.; Wu, J.; Yu, P. S.; and He, L. 2022. From known to unknown: Quality-aware self-improving graph neural network for open set social event detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 1696–1705.

Ren, J.; Peng, H.; Jiang, L.; Hao, Z.; Wu, J.; Gao, S.; Yu, Z.; and Yang, Q. 2024. Towards Cross-lingual Social Event Detection with Hybrid Knowledge Distillation. *ACM Transactions on Knowledge Discovery from Data*.

Ren, J.; Peng, H.; Jiang, L.; Wu, J.; Tong, Y.; Wang, L.; Bai, X.; Wang, B.; and Yang, Q. 2021. Transferring knowledge distillation for multilingual social event detection. *arXiv preprint arXiv:2108.03084*.

Sihem Sahnoun, S. E.; and Yahia, S. B. 2020. Event detection based on open information extraction and ontology. *Journal of Information and Telecommunication*, 4(3): 383–403.

Singh, T.; and Kumari, M. 2021. Burst: real-time events burst detection in social text stream. *The Journal of Supercomputing*, 77(10): 11228–11256.

Sun, L.; Huang, Z.; Peng, H.; Wang, Y.; Liu, C.; and Yu, P. S. 2024. LSEnet: Lorentz Structural Entropy Neural Network for Deep Graph Clustering. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.

Xue, F.; Hong, R.; He, X.; Wang, J.; Qian, S.; and Xu, C. 2019. Knowledge-based topic model for multi-modal social event analysis. *IEEE Transactions on Multimedia*, 22(8): 2098–2110.

Yang, Z.; Wei, Y.; Li, H.; Li, Q.; Jiang, L.; Sun, L.; Yu, X.; Hu, C.; and Peng, H. 2024. Adaptive Differentially Private Structural Entropy Minimization for Unsupervised Social Event Detection. *arXiv preprint arXiv:2407.18274*.

Yu, C.; Hu, B.; and Wang, Z. 2025. Open-world disaster information identification from multimodal social media. *Complex & Intelligent Systems*, 11(1): 7.

Yu, X.; Ren, J.; Jiang, L.; Peng, H.; Hao, Z.; Sun, L.; Peng, K.; Zhu, L.; and Yu, P. S. 2025a. PromptSED: An evolving topic-enhanced prompting framework for incremental social event detection. *Neural Networks*, 107772.

Yu, X.; Wei, Y.; Li, P.; Zhou, S.; Peng, H.; Sun, L.; Zhu, L.; and Yu, P. S. 2024. DAMe: Personalized Federated Social Event Detection with Dual Aggregation Mechanism. *arXiv preprint arXiv:2409.00614*.

Yu, X.; Wei, Y.; Zhou, S.; Yang, Z.; Sun, L.; Peng, H.; Zhu, L.; and Yu, P. S. 2025b. Towards effective, efficient and unsupervised social event detection in the hyperbolic space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 13106–13114.

Zhang, J.; Peng, H.; Sun, L.; Wu, G.; Liu, C.; and Yu, Z. 2025. Unsupervised graph clustering with deep structural entropy. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 3752–3763.

Zhang, K.; Yu, X.; Li, P.; Peng, H.; and Yu, P. S. 2024. SocialED: A Python Library for Social Event Detection. *arXiv preprint arXiv:2412.13472*.

Zhou, H.; Yin, H.; Zheng, H.; and Li, Y. 2020. A survey on multi-modal social event detection. *Knowledge-Based Systems*, 195: 105695.