

# Label-Free Contrastive Learning for Open-World Multimodal Social Event Detection

Zhiwei Yang

Institute of Information Engineering,  
Chinese Academy of Sciences  
Beijing, China  
School of Cyber Security, University  
of Chinese Academy of Sciences  
Beijing, China  
yangzhiwei@iie.ac.cn

Xiaoyan Yu

Beijing Institute of Technology  
Beijing, China  
xiaoyan.yu@bit.edu.cn

Haimei Qin\*

Institute of Information Engineering,  
Chinese Academy of Sciences  
Beijing, China  
School of Cyber Security, University  
of Chinese Academy of Sciences  
Beijing, China  
qinhaimei@iie.ac.cn

Li Sun

North China Electric Power  
University  
Beijing, China  
ccesunli@ncepu.edu.cn

Hao Peng

Beihang University  
Beijing, China  
penghao@buaa.edu.cn

Lei Jiang

Institute of Information Engineering,  
Chinese Academy of Sciences  
Beijing, China  
School of Cyber Security, University  
of Chinese Academy of Sciences  
Beijing, China  
jianglei@iie.ac.cn

## Abstract

Multimodal content on social media contains abundant cues about real-world events, and its automatic detection is critical for public safety and social governance. However, Multimodal Social Event Detection in the open world faces two major challenges: (1) They depend on supervised event labels or structured information; however, social media data in open-world settings often lack both, making it challenging for such methods to adapt to the dynamic nature of social media. (2) They rely on predefined label sets, i.e., the total number of events must generally be specified during the detection process. In contrast, in the open world, the total number of events is inherently difficult to estimate. To tackle these challenges, this paper proposes LFEEvent, a label-free contrastive learning framework for Multimodal Social Event Detection. To address the first challenge, we design a label-free multimodal contrastive learning strategy that relies solely on positive samples. Specifically, we design a multimodal large language model-based semantic enhancement strategy. Leveraging carefully crafted prompts, it enriches raw image-text pairs across three dimensions—event theme, event type, and image description—to construct robust positive samples. Subsequently, a dedicated Siamese Network enables self-supervised cross-modal alignment and representation learning. To address the second challenge, we introduce unsupervised clustering into the MSED task for the first time. A novel structure entropy-guided hierarchical clustering method is proposed, which automatically determines the number of event clusters and enables the detection

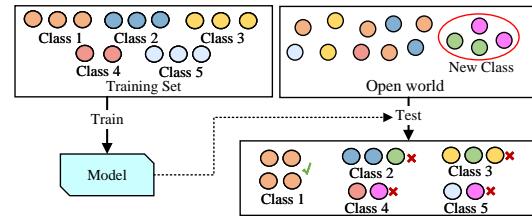


Figure 1: Supervised methods fail in the open world, as they cannot recognize new classes that appear during testing.

of unseen events in the training set. Experiments on multiple social media datasets demonstrate that LFEEvent significantly outperforms existing methods, especially in detecting previously unseen events.

## CCS Concepts

- Computing methodologies → Cluster analysis; Machine learning;
- Information systems → Clustering.

## Keywords

Multimodal Social Event Detection, Contrastive Learning, Structural Entropy

## ACM Reference Format:

Zhiwei Yang, Haimei Qin, Hao Peng, Xiaoyan Yu, Li Sun, and Lei Jiang. 2026. Label-Free Contrastive Learning for Open-World Multimodal Social Event Detection. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining (WSDM '26)*, February 22–26, 2026, Boise, ID, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3773966.3777919>

\*Corresponding author



This work is licensed under a Creative Commons Attribution 4.0 International License.  
WSDM '26, Boise, ID, USA

© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2292-9/2026/02  
<https://doi.org/10.1145/3773966.3777919>

## 1 INTRODUCTION

Social media platforms have become critical hubs for real-time public information exchange, continuously generating vast volumes

of multimodal user content encompassing text, images, and videos. Such content reflects authentic social events, including natural disasters, public gatherings, and cultural activities. Timely and accurate identification of these events is essential for understanding social dynamics, gauging public sentiment, and supporting crisis response, holding significant implications for domains such as sentiment analysis [3] and public safety assurance [24]. For instance, rapid detection of earthquake-related posts on Twitter can provide vital information for post-disaster relief. Consequently, the problem of open-world Multimodal Social Event Detection (MSED) is attracting increasing attention from both academia and industry [42].

MSED aims to automatically mine and identify events from massive heterogeneous social data, ultimately aggregating them into semantically consistent event clusters [44]. The heterogeneity of social data and the unpredictability of emerging events in open-world scenarios further intensify detection challenges. Current mainstream Social Event Detection (SED) and MSED methods predominantly rely on high-quality, structured data inputs, often requiring additional manual annotation. For example, supervised approaches [35] necessitate human labeling of event categories for training samples—a labor-intensive and costly task. Graph Neural Network (GNN)-based methods [4, 21, 29] depend on constructing graph structures using attributes such as user mentions, hashtags, user IDs, or keywords, yet such information is frequently unavailable in real-world scenarios, similarly demanding extra annotation or preprocessing. Thus, existing methods exhibit limitations when processing large-scale, dynamically evolving open social environments, hindering practical applicability. This constitutes our first core challenge: **how to eliminate reliance on predefined data formats and annotations to adapt to the open world**.

Furthermore, while some methods [1, 18] achieve promising performance in closed-world settings, they universally depend on predefined event labels and assume label consistency during testing. In open-world scenarios, however, novel events continuously emerge, rendering fixed label sets insufficient for comprehensive coverage and causing supervised methods to generalize poorly to unseen categories. As illustrated in Figure 1, the emergence of new events significantly degrades their performance. Recent studies [25, 26] attempt to enhance classifiers' capability to recognize unknown categories, partially adapting to open-world settings. Nevertheless, such methods can only group new events under a single "unknown" class, failing to distinguish distinct events into separate clusters—fundamentally persisting in task simplification. Thus, another critical question arises: **how to achieve fine-grained detection of unknown events without predefined labels**.

Facing these challenges, we propose a novel framework: Label-Free Contrastive Learning for Open-World Multimodal Social Event Detection (LFEEvent). This work pioneers Label-Free multimodal contrastive learning for open-world MSED, aiming to resolve both aforementioned challenges. **For the first challenge**, we design a *Label-Free Multimodal Contrastive Learning* strategy that relies solely on positive samples, eliminating labels or predefined categories to enable self-supervised learning directly on raw data, thereby enhancing representation capability. Specifically, we design a Multimodal Large Language Model (MLLM)-based Semantic Enhancement strategy to generate three granularity-enhanced textual descriptions (event theme, event type, and image caption) from

raw image-text pairs via carefully designed prompts. These enhanced texts, obtained through Visual Question Answering (VQA), construct robust positive samples. We then introduce a Siamese Network supporting multimodal fusion (with weight-sharing branches), where one branch processes raw image-text pairs and the other handles enhanced positive samples. Finally, self-supervised contrastive learning is performed to derive robust multimodal joint representations. **For the second challenge**, we formulate multimodal event detection as a clustering task to achieve fine-grained detection of novel events. We propose a novel *Structure Entropy (SE)-Guided Hierarchical Clustering* method that maps layers of a clustering tree into encoding trees. By computing the structural information of encoding trees, it automatically determines the optimal number of event clusters and guides the clustering process. This approach requires no predefined event count, thus better adapting to the dynamic emergence of new events in open-world scenarios. To our knowledge, while existing deep learning methods predominantly adopt classification paradigms, this work innovatively explores a clustering paradigm.

Our contributions are summarized as follows: 1) We propose LFEEvent, the first label-free multimodal contrastive learning framework for open-world MSED, eliminating reliance on labels or predefined categories. 2) We design a semantic enhancement strategy based on MLLMs and introduce a Siamese Network to achieve robust multimodal event representations. 3) We innovatively adopt a clustering paradigm, proposing a SE-guided hierarchical clustering strategy to enable automatic event discovery without prior knowledge of event quantity. 4) Extensive experiments on real-world social datasets demonstrate significant superiority over state-of-the-art methods, particularly in novel event recognition.

## 2 RELATED WORK

### 2.1 Unimodal Social Event Detection

Social event detection has long been a focal point of research, gradually evolving from unimodal to multimodal approaches. In the text-only setting, early methods modeled topics based on content [13, 32, 34, 39], and then applied clustering to detect social events. However, these approaches are strictly constrained by the limitations of topic modeling and often suffer from poor robustness. Currently, mainstream textual methods typically leverage GNNs for representation learning [4, 10, 21, 22, 29–31, 37, 40, 41, 43]. These methods take full advantage of the message-passing capabilities of GNNs to effectively represent events, followed by clustering for SED, achieving notable performance. However, their applicability is limited when essential metadata for graph construction is missing from the dataset. In the image-only setting, most methods utilize image posts and perform SED based on the publisher's posting time, geographic location, tags, and title [23, 36].

Nevertheless, the main limitation of unimodal methods lies in their reliance on a single type of data, which often leads to information loss and makes it difficult to capture a complete and detailed picture of events. This is particularly problematic in social media environments, where multimodal data is pervasive; relying on a single modality undermines detection performance.

## 2.2 Multimodal Social Event Detection

In recent years, some researchers have begun exploring multimodal social event detection. For example, KGE-MMSLDA [35] is a multimodal topic model that incorporates extensive external knowledge to perform knowledge-enhanced event classification. SCBD [1] is a multimodal event classification model in a closed-world setting that uses cross-modal attention to fuse image and text features. MFEK [18] integrates external knowledge and uses joint attention mechanisms to fuse text, images, and knowledge while filtering irrelevant information, achieving strong closed-world classification performance. Despite their successes, these approaches benefit from simplifying the MSED task into a classification problem. In an open-world setting, where a large number of novel events constantly emerge, models built on a fixed set of class labels are no longer applicable. OWSEC [25] proposes a masked transformer for modality fusion and designs an open-world classifier capable of detecting unseen categories. ODII [38] achieves disaster information identification in an open-world setting by designing a multi-task classifier. However, OWSEC and ODII can only group all novel events into the "unknown" class, failing to produce fine-grained event clusters.

Therefore, performing MSED using a clustering paradigm is a more appropriate and promising direction to handle dynamic event emergence in open-world scenarios. Unfortunately, due to the inherent challenges in learning robust representations from multimodal data, very few studies have explored this avenue.

## 3 PRELIMINARY

### 3.1 Task Definition

We provide the formal definition of MSED. Given a series of social messages  $P = \{p_1, \dots, p_n\}$ , where  $p_n = \{text_n, image_n\}$ , are text-image pairs. The task of MSED requires partitioning  $P$  into several partitions  $E = \{e_1, \dots, e_m\}$ , where  $e_i \cap e_j = \emptyset$ ,  $e_1 \cup \dots \cup e_m = P$ . Each partition  $e_i$  represents a distinct social event. In open-world settings, novel, previously unseen events often appear in the test set.

### 3.2 Encoding Tree and Structural Entropy

SE [14, 15] is a graph structure information measurement indicator based on an encoding tree. SE has a good effect in measuring clustering quality [33].

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

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

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.

## 4 METHODOLOGY

In this section, we provide a detailed description of LFEEvent. The entire framework is illustrated in Figure 2. LFEEvent operates in two stages. In the first stage, it introduces a novel self-supervised label-free contrastive learning strategy to learn robust multimodal representations of social events. In the second stage, it applies SE-guided hierarchical clustering to perform MSED without the need to predefine the number of clusters.

### 4.1 Label-Free Contrastive Learning

Robust message representation is essential for accurate MSED [5]. To improve representational capacity—particularly for emerging events in open-world settings—we propose a self-supervised label-free contrastive learning approach. Using only positive samples for contrastive learning has been proven to be feasible, such as BYOL [9] and SimSiam [6], but they only support single-modal contrastive learning for images. We have innovatively extended this to multimodal contrastive learning.

**4.1.1 MLLM-based Semantic Enhancement Strategy.** To construct positive samples for contrastive learning, we design an MLLM-based Semantic Enhancement strategy. VQA is one of the most representative tasks in MLLMs. VQA can augment multiple downstream tasks by mining potential knowledge in multimodal data and obtaining richer textual representations through pairs of question-and-answer formats [11, 17]. Our goal is to enhance the raw data to obtain higher-order representations of events. We use MLLMs to enhance the event. We make enhancements in three aspects: event type ( $E_{type}$ ), event theme ( $E_{theme}$ ), and image caption ( $E_{caption}$ ). Prompts are shown in Figure 2.

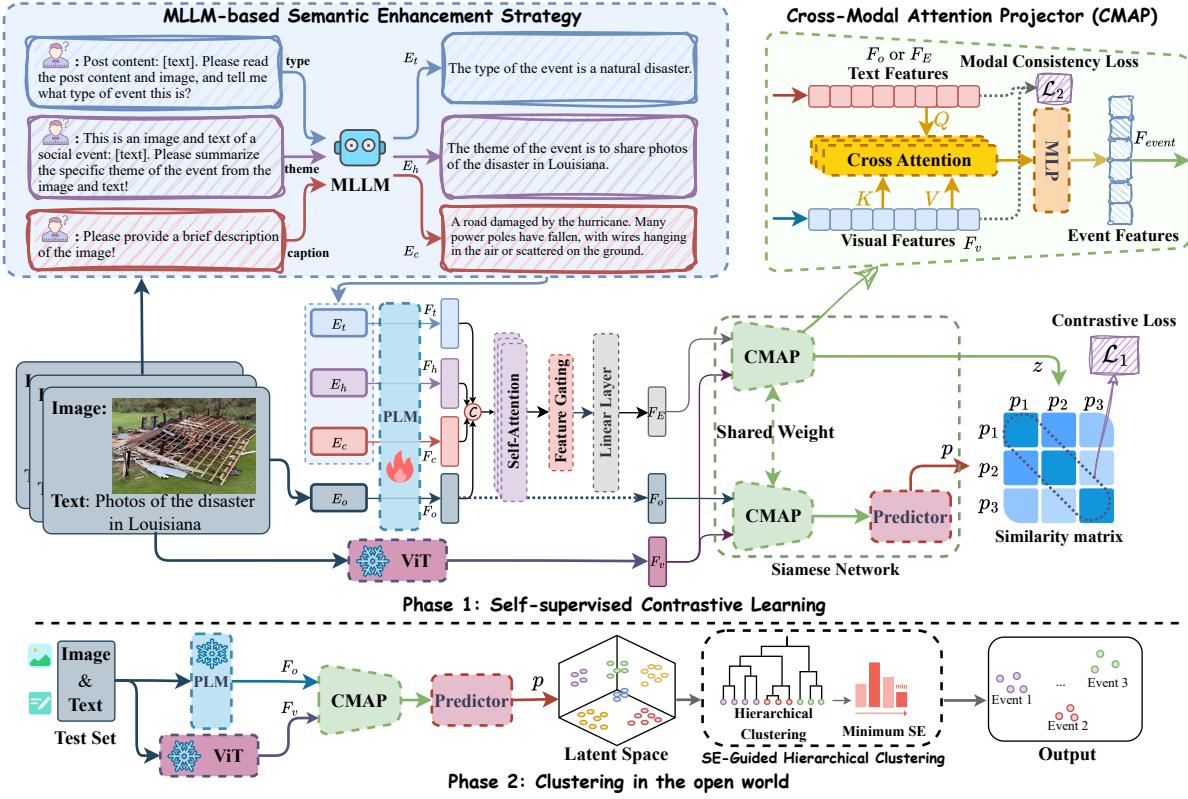
For  $E_{type}$ , it can roughly understand events such as natural disasters, financial crises, and terrorist incidents etc. For  $E_{theme}$ , it can provide a more detailed understanding of the event, including specific themes or topics within the event type, such as "hurricane consequences", "market crash impact", or "terrorist attack details", which will help us judge the specific event. For  $E_{caption}$ , it can generate descriptive captions for the images, which help in understanding the visual content of the posts.

**4.1.2 Representation of Positive Sample.** To obtain a unified text representation for positive samples, we dynamically integrate multi-source text features via self-attention mechanisms and gating mechanisms.

First, we utilize a Pretrained Language Model (PLM) to obtain feature embeddings for the three types of augmented text and the original text:

$$\mathbf{F}_o, \mathbf{F}_t, \mathbf{F}_h, \mathbf{F}_c = \text{PLM}(\mathbf{E}_o, \mathbf{E}_t, \mathbf{E}_h, \mathbf{E}_c), \quad (2)$$

where  $\mathbf{E}_o$  is raw text. Subsequently, the aforementioned four types of embeddings are stacked in the feature dimension, and a Multi-Head Self-Attention is employed to model interactions for the



**Figure 2: The proposed LFEVENT framework.** The first phase illustrates the model training process: enhanced texts are first obtained via the MLLM-based semantic enhancement strategy; subsequently, positive sample representations are extracted through attention and gating mechanisms; finally, contrastive learning is performed using a Siamese Network. The second phase demonstrates how MSED is conducted in the open world: event representations are first derived, and event clusters are then obtained via the SE-guided hierarchical clustering algorithm.

stacked features, capturing high-order dependencies among different text sources.

$$[\mathbf{a}_o, \mathbf{a}_t, \mathbf{a}_h, \mathbf{a}_c] = \text{MultiHeadAttn}([\mathbf{F}_o, \mathbf{F}_t, \mathbf{F}_h, \mathbf{F}_c]). \quad (3)$$

Next, an independent gating network is designed for each type of text feature, with fusion weights obtained through Sigmoid activation. The gating weights can be dynamically adjusted based on the actual input, reflecting the importance of each text type for the current sample.

$$\begin{aligned} [g_o, g_t, g_h, g_c] &= \text{Softmax}([\text{Gate}_o(\mathbf{F}_o), \\ &\text{Gate}_t(\mathbf{F}_t), \text{Gate}_h(\mathbf{F}_h), \text{Gate}_c(\mathbf{F}_c)]). \end{aligned} \quad (4)$$

Finally, the self-attention output is multiplied by the gating weights and concatenated, then fed into a fusion transformation layer (comprising two linear layers and a normalization layer), ultimately yielding text representations of positive samples  $\mathbf{F}_E$ .

$$\mathbf{F}_E = \text{Linear}([g_o \cdot \mathbf{a}_o, g_t \cdot \mathbf{a}_t, g_h \cdot \mathbf{a}_h, g_c \cdot \mathbf{a}_c]). \quad (5)$$

The final representation of the original sample is  $\mathbf{F}_o$ .

The visual representations are extracted using a frozen pre-trained Visual Transformer (ViT) [8]. The visual representation is expressed as:

$$\mathbf{F}_v = \text{ViT}(img). \quad (6)$$

**4.1.3 Siamese Network of LFEVENT.** The backbone network structure of LFEVENT is a Siamese Network. We designed the Cross-Modal Attention Projector (CMAP) to fuse multi-modal features, and CMAP shares weights in the Siamese Network.

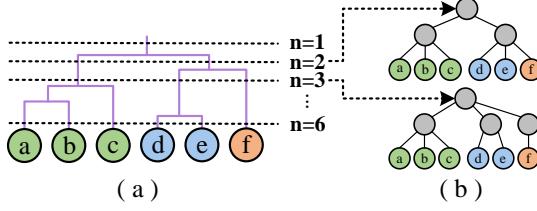
To achieve a deep fusion of image and text features to represent events, we design the CMAP module, as shown in the green box in Figure 2. The text feature input to CMAP is  $\mathbf{F}_E$  or  $\mathbf{F}_o$ . Subsequently, we use the cross-attention to fuse the text and image features. We believe that text information is the dominant component in a message, while image information is auxiliary; therefore, we utilized text-to-image cross-attention. Take  $\mathbf{F}_E$  as an example, the attention score output  $S'$  is defined as:

$$S' = \text{softmax}\left(\frac{(\mathbf{F}_E \mathbf{W}_Q)(\mathbf{F}_v \mathbf{W}_K)^T}{\sqrt{d_k}}\right)(\mathbf{F}_v \mathbf{W}_V), \quad (7)$$

where  $d_k$  is the same as the dimension of  $\mathbf{F}_E$ .  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$ , and  $\mathbf{W}_V$  are the projection weight matrices to be learned. Ultimately, the module's output is the fused feature  $\mathbf{F}_{event}$ :

$$\mathbf{F}_{event} = S' \mathbf{W}_O, \quad (8)$$

where  $\mathbf{W}_O$  are the projection weight matrices to be learned. Notably, we consider image features as a complement and auxiliary to text



**Figure 3:** (a): A toy example of the tree of clusters. (b): The process of mapping the tree of clusters to the encoding tree.

features. Therefore, we freeze the ViT and adopt a unidirectional cross-attention mechanism, which saves computational resources and facilitates scalability.

As illustrated in Figure 2, one branch of the Siamese Network comprises the CMAP modules to process enhanced features. The other branch, consisting of the CMAP and a Predictor, processes features from the original data. The Predictor is a simple linear model, which can mitigate training collapse. While both branches share a similar structure, they differ in the type of input data they use. During contrastive learning, the weights of both branches are alternately updated, enabling effective representation learning.

**4.1.4 Training Objectives.** In addition to the contrastive learning loss, we introduce an auxiliary modal consistency loss to enhance training stability. We adopt a two-part loss function consisting of a contrastive loss and a modal consistency loss. Contrastive loss is expressed as:

$$\mathcal{L}_1 = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\langle \mathbf{p}_i, \mathbf{z}_i \rangle / \tau)}{\sum_{j=1}^N \exp(\langle \mathbf{p}_i, \mathbf{z}_j \rangle / \tau)}, \quad (9)$$

and the modal consistency loss is expressed as:

$$\mathcal{L}_2 = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\langle \mathbf{v}_i, \mathbf{t}_i \rangle / \tau)}{\sum_{j=1}^N \exp(\langle \mathbf{v}_i, \mathbf{t}_j \rangle / \tau)}, \quad (10)$$

so total loss is expressed as:

$$\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_2, \quad (11)$$

where  $\mathbf{p}_i$  and  $\mathbf{z}_i$  represent the normalized predicted features and target features, respectively, derived from the  $F_{event}$  outputs of both sides of the Siamese Network.  $\mathbf{v}_i$  and  $\mathbf{t}_i$  represent the normalized image and text features.  $\langle \cdot, \cdot \rangle$  denotes the dot product similarity,  $\tau$  is the temperature parameter,  $\lambda$  is the weighting factor for  $\mathcal{L}_2$  and  $N$  is the batch size.

## 4.2 Clustering in the Open World

During detection, only the ViT, PLM, CMAP, and Predictor modules are needed to complete the detection, while the more computationally intensive MLLM-based Semantic Enhancement module is discarded, designed to accommodate large-scale social media data.

**4.2.1 SE-Guided Hierarchical Clustering.** To detect events in the latent space without predefining the number of events, we propose an SE-Guided Hierarchical Clustering algorithm. Hierarchical clustering [20], a well-established machine learning technique, constructs a dendrogram  $\mathcal{T}_c$  (illustrated in Figure 3, (a)) by iteratively merging

---

**Algorithm 1:** Structural Entropy-Guided Hierarchical Clustering

---

**Input:**  $\mathcal{T}_c$  with  $N$  leaf nodes; number of nearest neighbors  $k$ .  
**Output:** Final clustering result  $E$ .

```

1 Initialize  $V$  as the set of all leaf nodes in  $\mathcal{T}_c$ ;
2 Initialize  $E \leftarrow \emptyset$  as the set of edges;
3 Initialize  $\mathcal{G} = (V, E)$ ;
4 for each node  $v_i \in V$  do
5   Compute cosine similarity between  $v_i$  and all other
      nodes in  $V$ ;
6   Identify the  $k$  nodes with the highest cosine similarity
      scores to  $v_i$ ;
7   Add edges connecting  $v_i$  to these  $k$  nodes in the graph;
8 Initialize structural entropy sequence  $SE \leftarrow \emptyset$ ;
9 Set  $n \leftarrow N$  as the initial number of clusters;
10 while  $n > 0$  do
11   Map the current tree structure  $\mathcal{T}_c$  to an encoding tree  $\mathcal{T}_n$ 
      with  $n$  clusters;
12   Compute structural entropy  $se_n$  of  $\mathcal{T}_n$  using Equation 1;
13   Append  $se_n$  to the sequence  $SE$ ;
14   Decrement  $n \leftarrow n - 1$ ;
15 Determine  $n_{best} \leftarrow \arg \min(SE)$  as the number of clusters
      with minimum structural entropy;
16 Obtain the clustering result  $E = \{e_1, e_2, \dots, e_{n_{best}}\}$ 
      corresponding to  $n_{best}$ ;
17 return  $E$ ;

```

---

smaller clusters into larger ones. However, traditional hierarchical clustering requires a predetermined number of clusters, which is impractical for MSED in open-world settings. To address this, we map each layer of the hierarchical clustering tree to an encoding tree of height 2 to compute the two-dimensional SE. The optimal number of clusters  $n$  is automatically determined by minimizing the SE. Specifically, given a dendrogram  $\mathcal{T}_c$  with  $N$  leaf nodes (representing  $N$  event messages), we map the  $n$  clusters associated with a given cluster count  $n$  to the nodes  $\alpha$  of an encoding tree  $\mathcal{T}$  (as shown in Figure 3, (b)). Ultimately,  $\mathcal{T}_c$  is mapped to  $N$  encoding trees  $\mathcal{T}$ .

The detailed procedure is outlined in Algorithm 1. We utilize scikit-learn's hierarchical clustering algorithm to generate the cluster tree  $\mathcal{T}_c$  (Input). Initially, we construct a graph  $\mathcal{G} = (V, E)$  using the  $k$ -nearest neighbor method, with cosine similarity as the distance metric (lines 1–7). Next, we map each layer of  $\mathcal{T}_c$  to an encoding tree  $\mathcal{T}_n$  and compute the two-dimensional structural entropy (SE) (lines 10–14). We then identify the optimal number of clusters  $n_{best}$  corresponding to the minimum SE (line 15). Finally, the clustering result for  $n_{best}$  serves as the final event detection output (lines 16–17).

Algorithm 1 does not require specifying the number of clusters or any manual labeling, making it more competitive in the open world.

**Table 1: (Q1) The results of all methods in the open-world setting. (Bold indicates the best result, and underlined indicates the second-best result. The variance of the 5 runs follows “ $\pm$ ”. \* marks results acquired with the ground truth event numbers.**

Methods	NED (S1)			CrisisMMD (S1)			NED (S2)			CrisisMMD (S2)		
	ARI	NMI	AMI									
BERT*	.12 $\pm$ .00	.33 $\pm$ .00	.28 $\pm$ .00	.08 $\pm$ .00	.14 $\pm$ .00	.14 $\pm$ .00	.11 $\pm$ .00	.30 $\pm$ .00	.27 $\pm$ .00	.05 $\pm$ .00	.10 $\pm$ .00	.09 $\pm$ .00
SBERT*	.23 $\pm$ .00	.53 $\pm$ .00	.50 $\pm$ .00	.08 $\pm$ .00	.44 $\pm$ .00	.43 $\pm$ .00	.29 $\pm$ .00	.54 $\pm$ .00	.52 $\pm$ .00	.10 $\pm$ .00	.47 $\pm$ .00	.47 $\pm$ .00
CLIP (text)*	.28 $\pm$ .00	.56 $\pm$ .00	.53 $\pm$ .00	.09 $\pm$ .00	.42 $\pm$ .00	.41 $\pm$ .00	.26 $\pm$ .00	.54 $\pm$ .00	.51 $\pm$ .00	.12 $\pm$ .00	.47 $\pm$ .00	.47 $\pm$ .00
CLIP (image)*	.19 $\pm$ .00	.51 $\pm$ .00	.48 $\pm$ .00	.04 $\pm$ .00	.17 $\pm$ .00	.16 $\pm$ .00	.23 $\pm$ .00	.48 $\pm$ .00	.46 $\pm$ .00	.04 $\pm$ .00	.22 $\pm$ .00	.21 $\pm$ .00
HISEvent	.19 $\pm$ .00	.45 $\pm$ .00	.42 $\pm$ .00	.15 $\pm$ .00	.39 $\pm$ .00	.38 $\pm$ .00	.17 $\pm$ .00	.40 $\pm$ .00	.39 $\pm$ .00	.17 $\pm$ .00	.38 $\pm$ .00	.39 $\pm$ .00
BLIP2*	.11 $\pm$ .00	.29 $\pm$ .00	.26 $\pm$ .00	.16 $\pm$ .00	.42 $\pm$ .00	.41 $\pm$ .00	.25 $\pm$ .00	.39 $\pm$ .00	.41 $\pm$ .00	.16 $\pm$ .00	.40 $\pm$ .00	.40 $\pm$ .00
LLaVA*	.29 $\pm$ .00	.54 $\pm$ .00	.56 $\pm$ .00	.17 $\pm$ .00	.43 $\pm$ .00	.45 $\pm$ .00	.30 $\pm$ .00	.41 $\pm$ .00	.41 $\pm$ .00	.15 $\pm$ .00	.45 $\pm$ .00	.44 $\pm$ .00
BERT (caption)*	.15 $\pm$ .00	.38 $\pm$ .00	.34 $\pm$ .00	.02 $\pm$ .00	.06 $\pm$ .00	.06 $\pm$ .00	.14 $\pm$ .00	.36 $\pm$ .00	.33 $\pm$ .00	.05 $\pm$ .00	.14 $\pm$ .00	.14 $\pm$ .00
SBERT (caption)*	.27 $\pm$ .00	.57 $\pm$ .00	.54 $\pm$ .00	.10 $\pm$ .00	.47 $\pm$ .00	.46 $\pm$ .00	.28 $\pm$ .00	.56 $\pm$ .00	.54 $\pm$ .00	.11 $\pm$ .00	.49 $\pm$ .00	.49 $\pm$ .00
HISEvent (caption)	.22 $\pm$ .00	.52 $\pm$ .00	.50 $\pm$ .00	.17 $\pm$ .00	.40 $\pm$ .00	.41 $\pm$ .00	.19 $\pm$ .00	.45 $\pm$ .00	.48 $\pm$ .00	.16 $\pm$ .00	.40 $\pm$ .00	.39 $\pm$ .00
CLIP*	.32 $\pm$ .00	.61 $\pm$ .00	.59 $\pm$ .00	.04 $\pm$ .00	.21 $\pm$ .00	.20 $\pm$ .00	.33 $\pm$ .00	.64 $\pm$ .00	.59 $\pm$ .00	.05 $\pm$ .00	.27 $\pm$ .00	.26 $\pm$ .00
MMBT*	.21 $\pm$ .06	.45 $\pm$ .07	.45 $\pm$ .06	.19 $\pm$ .01	.25 $\pm$ .12	.25 $\pm$ .03	.23 $\pm$ .09	.50 $\pm$ .09	.51 $\pm$ .09	.15 $\pm$ .05	.28 $\pm$ .12	.28 $\pm$ .11
SCBD*	.30 $\pm$ .09	.56 $\pm$ .08	.55 $\pm$ .03	.25 $\pm$ .11	.39 $\pm$ .09	.40 $\pm$ .08	.26 $\pm$ .07	.46 $\pm$ .06	.43 $\pm$ .07	.19 $\pm$ .05	.24 $\pm$ .08	.24 $\pm$ .06
OWSEC*	.33 $\pm$ .02	.64 $\pm$ .01	.62 $\pm$ .02	.35 $\pm$ .02	.47 $\pm$ .01	.41 $\pm$ .01	.28 $\pm$ .01	.55 $\pm$ .01	.53 $\pm$ .03	.32 $\pm$ .02	.41 $\pm$ .01	.40 $\pm$ .02
MFEK*	.34 $\pm$ .04	.66 $\pm$ .03	.67 $\pm$ .02	.34 $\pm$ .04	.45 $\pm$ .05	.43 $\pm$ .03	.29 $\pm$ .06	.59 $\pm$ .04	.58 $\pm$ .06	.30 $\pm$ .04	.42 $\pm$ .04	.40 $\pm$ .03
ODII*	.30 $\pm$ .03	.55 $\pm$ .05	.55 $\pm$ .05	.24 $\pm$ .10	.29 $\pm$ .09	.30 $\pm$ .03	.27 $\pm$ .05	.59 $\pm$ .07	.57 $\pm$ .06	.19 $\pm$ .06	.26 $\pm$ .05	.27 $\pm$ .03
LFEVENT	<b>.90</b> $\pm$ .01	<b>.89</b> $\pm$ .02	<b>.89</b> $\pm$ .02	<b>.94</b> $\pm$ .03	<b>.93</b> $\pm$ .02	<b>.93</b> $\pm$ .03	<b>.85</b> $\pm$ .03	<b>.87</b> $\pm$ .03	<b>.86</b> $\pm$ .04	<b>.62</b> $\pm$ .01	<b>.82</b> $\pm$ .02	<b>.82</b> $\pm$ .02
Improve (%)	↑164	↑35	↑33	↑169	↑98	↑102	↑158	↑36	↑46	↑93	↑67	↑67

4.2.2 *Time Complexity.* In previous work, HISEvent used a greedy 2D SE minimization algorithm that searched for the optimal encoding tree by moving leaf nodes, resulting in high time complexity. We used a tree-based mapping 2D SE minimization algorithm that can complete clustering in a smaller iteration set, effectively reducing time complexity.

Specifically, the time complexity comprises  $O(n^2)$  for hierarchical clustering with the ‘ward’ method,  $O(n)$  for mapping to the encoding tree, and  $O(m + n)$  to  $O(mn)$  for computing 2D SE, the latter increasing with the number of graph communities. The total complexity is their sum. The contrastive learning and distillation modules, leveraging fine-tuning with most parameters frozen, each have a time complexity of  $O(n^2)$ . Thus, LFEVENT efficiently processes large-scale social media data streams.

## 5 EXPERIMENTS

In this section, our goal is to answer the following questions:

Q1: How does LFEVENT perform compared to other baselines on MSED?

Q2: How do the various parts of LFEVENT contribute to results?

Q3: How robust are different MLLMs?

Q4: What effect do the parameters of LFEVENT have on the performance of clustering?

Q5: How effective is LFEVENT in visual display?

### 5.1 Experimental Setup

5.1.1 *Datasets.* We evaluate on two datasets. The **NED** dataset [18] contains 17,366 image-text pairs from Twitter, annotated with 40 real-world events. The **CrisisMMD** dataset [2] consists of 16,097 image-tweet pairs covering seven natural disasters in 2017.

For the open-world settings, we simulate two splitting settings (S1 and S2) via stratified sampling. For **NED**, S1 uses 30 events for training/validation/test (80%/10%/10%), with the remaining 10 events split equally between validation and test to simulate unseen events. S2 samples from 20 events, with the remaining 20 treated as new events, maintaining the same proportions. For **CrisisMMD**, S1 uses five events for training and two as unseen; S2 uses four for training and three as unseen, following the same ratio as **NED**. For the closed-world settings, we use stratified sampling for training/validation/test (70%/10%/20%).

5.1.2 *Baselines.* For unimodal baselines, **BERT** [7] and **SBERT** [28] are used to extract text features, followed by K-means clustering. **CLIP** [27] extract text or image features, followed by K-means clustering. **HISEVENT** [5], the SOTA method in text modal.

For multimodal baselines, we generate image captions and concatenate them with the corresponding text to construct multimodal inputs. Based on this, we extend the following models as multimodal baselines: **HISEVENT**, **BERT**, **SBERT**, we adopt two pre-trained MLLMs, **BLIP2** [16] and **LLaVA** [19]. We extract features from their multimodal fusion layers and apply K-means clustering for evaluation. For classification models, we perform clustering on their embeddings: **MMBT** [12] employs a transformer to fuse textual and visual features for enhanced classification. **SCBD** [1] assigns weights to embeddings to emphasize important features across different modalities. **OWSEC** [25] is an open-world social event classification method that utilizes a masked transformer. **MFEK** [18] is a closed-world social event classification method based on external knowledge. **ODII** [38] achieves disaster information identification in an open-world setting by designing a multi-task classifier.

**Table 2: (Q1) The results of all methods in the closed-world setting. (Bold indicates the best result, and underlined indicates the second-best result. The variance of the 5 runs follows “ $\pm$ ”.)**

Datasets	Metrics	Methods									
		BERT	SBERT	HISEvent	CLIP	MMBT	SCBD	OWSEC	MFEK	ODII	LFEEvent
NED	ARI	.12 $\pm$ .00	.25 $\pm$ .00	.25 $\pm$ .00	<b>.70</b> $\pm$ .00	.55 $\pm$ .09	.69 $\pm$ .07	<u>.73</u> $\pm$ .05	<u>.72</u> $\pm$ .04	.68 $\pm$ .09	.72 $\pm$ .04
	NMI	.19 $\pm$ .00	.33 $\pm$ .00	.51 $\pm$ .00	.72 $\pm$ .00	.57 $\pm$ .12	.71 $\pm$ .06	<b>.82</b> $\pm$ .08	.75 $\pm$ .05	.70 $\pm$ .09	.82 $\pm$ .06
	AMI	.20 $\pm$ .00	.35 $\pm$ .00	.50 $\pm$ .00	.72 $\pm$ .00	.58 $\pm$ .11	.71 $\pm$ .09	<u>.82</u> $\pm$ .09	.76 $\pm$ .06	.70 $\pm$ .09	.83 $\pm$ .09
CrisisMMD	ARI	.03 $\pm$ .00	.59 $\pm$ .00	.72 $\pm$ .00	<u>.79</u> $\pm$ .00	.58 $\pm$ .07	.73 $\pm$ .12	.74 $\pm$ .09	.72 $\pm$ .02	.75 $\pm$ .07	.80 $\pm$ .10
	NMI	.07 $\pm$ .00	.67 $\pm$ .00	.76 $\pm$ .00	.76 $\pm$ .00	.57 $\pm$ .09	.78 $\pm$ .11	<u>.80</u> $\pm$ .10	.79 $\pm$ .05	.75 $\pm$ .08	.81 $\pm$ .08
	AMI	.07 $\pm$ .00	.66 $\pm$ .00	.75 $\pm$ .00	.75 $\pm$ .00	.57 $\pm$ .09	.78 $\pm$ .10	<u>.80</u> $\pm$ .08	.79 $\pm$ .08	.76 $\pm$ .09	.81 $\pm$ .06

**5.1.3 Implementation Details.** For the baseline using K-means, we set the number of true events instead of the number of events in the training set. Although this is unfair to our LFEEvent, our method still achieves better results. For other baselines, we use the original paper parameters. For LFEEvent, we set the learning rate to 0.05, and during training, the number of epochs is set to 30, with  $k$  set to 3,  $\tau$  to 0.07, and  $\lambda$  to 0.3. In the MLLM-based semantic enhancement strategy, we utilize the Qwen2-VL model for semantic enhancement. The PLM is SBERT. The ViT is the visual component of CLIP. We use 8 RTX 3090 GPUs. We report the mean over 5 runs for all experiments.

**5.1.4 Evaluation Metrics.** We learn from existing work [29] and measure adjusted mutual information (**AMI**), normalized mutual information (**NMI**), and adjusted rand index (**ARI**).

## 5.2 Overall Performance (Q1)

**5.2.1 Open-World Setting.** Table 1 presents the performance comparison between LFEEvent and all baseline methods under the open-world setting. Across both datasets and evaluation splits, LFEEvent ranks first on all three metrics, with clear advantages over the strongest competitors.

On the NED dataset, LFEEvent achieves substantial improvements over competitive multimodal methods such as MFEK, CLIP, and OWSEC. Specifically, in S1, LFEEvent outperforms the best baseline MFEK by **164%** in ARI, **35%** in NMI, and **33%** in AMI; in S2, it surpasses the second-best CLIP by **158%**, **36%**, and **46%** on the same metrics. While S2 shows marginally lower performance than S1 due to the higher proportion of unseen events, LFEEvent’s label-free paradigm facilitates self-supervised contrastive learning without manual annotations, maintaining robustness in dynamic and continually evolving environments. On the CrisisMMD dataset, LFEEvent demonstrates even greater relative gains over strong baselines such as SCBD, MFEK, and OWSEC. For example, in S1, compared to the best baseline, LFEEvent achieves improvements of **169%** in ARI, **98%** in NMI, and **102%** in AMI; in S2, the gains remain substantial at **93%**, **67%**, and **67%**. These results further reveal the limitations of supervised learning paradigms in open-world environments, where their dependence on fixed label sets severely weakens their ability to distinguish newly emerging events.

Overall, the experiments lead to two consistent observations:

(1) Multimodal methods significantly outperform unimodal ones, validating the necessity of integrating textual and visual modalities in open-world social event detection.

(2) Owing to its Label-Free Contrastive Learning strategy, LFEEvent effectively learns fine-grained event semantics from noisy multimodal social media streams, achieving superior clustering performance even with a large proportion of unseen events, and without the need to predefine the number of clusters.

**5.2.2 Closed-World Setting.** Table 2 presents the performance under the Closed-World Setting. In conducting the closed-world setting experiments, the baseline models were trained in a classification manner, after which both classification labels and clustering labels were obtained, and the best results were reported.

LFEEvent demonstrates strong performance on both datasets. Among unimodal baselines, HISEvent achieves the best results, benefiting from the robustness of structural entropy. However, multimodal methods generally outperform unimodal ones, highlighting the importance of image information in MSED. Within the multimodal baselines, given the closed-world restriction, these methods can perform relatively well. CLIP, leveraging its powerful alignment and feature representation capabilities, achieves promising performance. MMBT, SCBD, and ODII rely on feature fusion and supervision, while MFEK incorporates extensive external knowledge, thus delivering better results. OWSEC employs a more advanced masked Transformer for fine-grained image–text feature fusion, thereby achieving the strongest baseline performance. Nevertheless, their success is based on the closed-world assumption—no new events emerge, and the predefined label set remains fixed. LFEEvent, free from the constraint of a predefined label set, still achieves encouraging results.

## 5.3 Ablation Study (Q2)

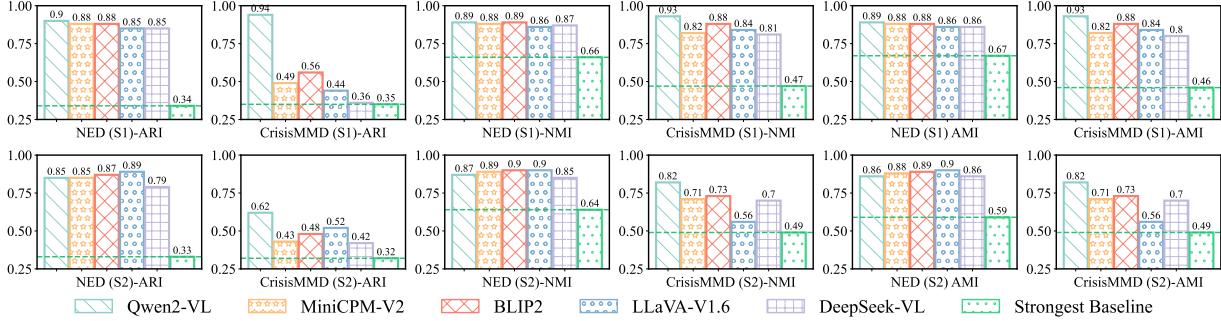
We perform an ablation study to evaluate the contribution of each component in LFEEvent, with results shown in Table 3. Each ablation removes a single module from the full model to assess its impact.

The removal of the modal consistency loss  $\mathcal{L}_2$  leads to a consistent and noticeable performance degradation across all datasets and settings. The most pronounced decline is observed on NED (S2), where the ARI drops by 0.11, NMI by 0.07, and AMI by 0.06 compared to the full model. This suggests that  $\mathcal{L}_2$  plays a critical role in aligning multimodal representations, ensuring robust feature consistency across diverse data modalities, particularly in challenging open-world scenarios.

Eliminating the event-type enhancement module  $E_{\text{type}}$  results in the most significant performance drop on CrisisMMD (S1), with ARI, NMI, and AMI decreasing by 0.11, 0.12, and 0.12, respectively.

**Table 3: (Q2) Ablation study results for different configurations.**

#	Configurations				NED (S1)			CrisisMMD (S1)			NED (S2)			CrisisMMD (S2)		
	$\mathcal{L}_2$	$E_{\text{type}}$	$E_{\text{theme}}$	$E_{\text{caption}}$	ARI	NMI	AMI									
1	✗	✓	✓	✓	0.85 <sub>.05</sub>	0.85 <sub>.04</sub>	0.84 <sub>.05</sub>	0.85 <sub>.09</sub>	0.85 <sub>.08</sub>	0.84 <sub>.09</sub>	0.74 <sub>.11</sub>	0.80 <sub>.07</sub>	0.80 <sub>.06</sub>	0.55 <sub>.07</sub>	0.75 <sub>.07</sub>	0.76 <sub>.06</sub>
2	✓	✗	✓	✓	0.75 <sub>.15</sub>	0.75 <sub>.14</sub>	0.74 <sub>.75</sub>	0.83 <sub>.11</sub>	0.81 <sub>.12</sub>	0.81 <sub>.12</sub>	0.79 <sub>.06</sub>	0.75 <sub>.12</sub>	0.75 <sub>.11</sub>	0.49 <sub>.13</sub>	0.68 <sub>.14</sub>	0.68 <sub>.14</sub>
3	✓	✓	✗	✓	0.78 <sub>.12</sub>	0.77 <sub>.12</sub>	0.77 <sub>.12</sub>	0.78 <sub>.16</sub>	0.77 <sub>.16</sub>	0.77 <sub>.16</sub>	0.70 <sub>.15</sub>	0.75 <sub>.12</sub>	0.75 <sub>.11</sub>	0.42 <sub>.20</sub>	0.69 <sub>.13</sub>	0.69 <sub>.13</sub>
4	✓	✓	✓	✗	0.85 <sub>.05</sub>	0.84 <sub>.05</sub>	0.84 <sub>.05</sub>	0.93 <sub>.01</sub>	0.91 <sub>.02</sub>	0.91 <sub>.02</sub>	0.79 <sub>.06</sub>	0.86 <sub>.01</sub>	0.85 <sub>.01</sub>	0.60 <sub>.02</sub>	0.79 <sub>.03</sub>	0.80 <sub>.02</sub>
5	✓	✓	✓	✓	0.90	0.89	0.89	0.94	0.93	0.93	0.85	0.87	0.86	0.62	0.82	0.82

**Figure 4: (Q3) Performance comparison of semantic enhancement using different MLLMs. The strongest baseline is the optimal result among all baselines.**

This substantial degradation highlights the critical importance of event-type cues in providing discriminative power for effective event separation, particularly in scenarios with well-defined event categories.

The exclusion of the event-theme enhancement module  $E_{\text{theme}}$  causes notable performance declines, with the most severe impact observed on CrisisMMD (S2), where ARI drops by 0.20, and both NMI and AMI decrease by 0.13. This underscores the role of thematic abstraction in capturing high-level semantic patterns, which enhances the model’s generalization capability, especially for evolving or dynamic events in open-world settings.

Removing the event-caption enhancement module  $E_{\text{caption}}$  results in moderate performance drops, with a more pronounced effect on NED (S2), where ARI decreases by 0.06, NMI by 0.01, and AMI by 0.01. While the contribution of caption-based cues is less dominant compared to event-type or theme enhancements, they still provide valuable contextual information, improving multimodal alignment and enriching the construction of positive samples for clustering.

Overall, each component of LFEEvent contributes to its robust performance, and their synergistic integration ensures both stability and effectiveness in open-world social event detection.

#### 5.4 Analysis of MLLMs(Q3)

To study the impact of different MLLMs on LFEEvent, we use five representative MLLMs for positive sample construction. Among them, MiniCPM-V2 contains 2.8 billion parameters, while the other models have 7 billion parameters. We evaluated their performance in an open-world setting, and the results are shown in Figure 4.

Across all datasets and metrics, Qwen2-VL achieves the highest or near-highest performance, indicating its superior capability in extracting informative event semantics from image–text pairs. On

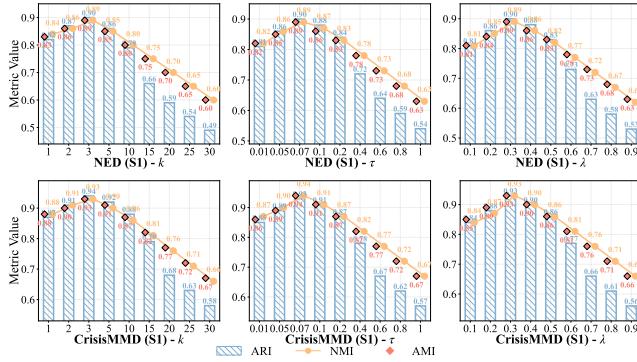
the NED dataset, Qwen2-VL consistently attains the best scores, reaching 0.90/0.89/0.89 (ARI/NMI/AMI) in S1 and 0.85/0.87/0.86 in S2. On CrisisMMD, Qwen2-VL also leads with large margins, especially in ARI, where it surpasses the strongest baseline by 0.59 in S1 and 0.30 in S2. MiniCPM-V2, BLIP2, and LLaVA-V1.6 deliver competitive results, with BLIP2 showing strong overall stability and LLaVA-V1.6 achieving notable gains in NMI and AMI on NED. DeepSeek-VL, while effective on NED, exhibits a larger drop on CrisisMMD, suggesting less robustness to domain shifts in event types and visual styles. The comparison with the strongest baseline (without semantic enhancement) highlights the necessity of the MLLM-based enhancement strategy. Even the lowest-performing MLLM configuration substantially outperforms the strongest baseline, demonstrating that enriching raw multimodal inputs with high-quality semantic cues is a key factor for LFEEvent’s success.

In summary, LFEEvent benefits consistently from different MLLMs, but the choice of MLLM can significantly influence the absolute performance. High-capacity and domain-adaptive MLLMs, such as Qwen2-VL, provide the most effective semantic enrichment, enabling LFEEvent to achieve superior clustering quality in open-world multimodal social event detection.

#### 5.5 Hyperparameter Sensitivity (Q4)

We evaluate the influence of three key parameters on LFEEvent: the number of neighbors  $k$  in Algorithm 1, the temperature  $\tau$  in the contrastive loss, and the weight  $\lambda$  of the modal consistency loss. Each parameter is varied independently while others are fixed, and results are reported in ARI, NMI, and AMI on NED (S1) and CrisisMMD (S1) (Figure 5).

For  $k$ , performance exhibits a clear unimodal trend: too small  $k$  leads to sparse graphs with insufficient structural information,



**Figure 5: (Q4) Performance of LFEEvent (Qwen2-VL) under different parameters.**

while overly large  $k$  introduces noise through redundant edges, degrading clustering quality. The optimal range  $k \in [1, 5]$  achieves the best balance, though larger values increase the cost of computing 2D SE. For  $\tau$ , extremely low values sharpen the similarity distribution excessively, amplifying noise, while high values over-smooth it, weakening feature discrimination. A moderate range  $\tau \in [0.05, 0.2]$  consistently yields high performance, indicating an effective trade-off between cluster compactness and separation. For  $\lambda$ , small values underweight the modal consistency constraint, allowing semantic drift between modalities; large values overemphasize cross-modal alignment, reducing intra-modal separability. The range  $\lambda \in [0.2, 0.5]$  offers optimal performance, underscoring the importance of balanced multi-modal supervision.

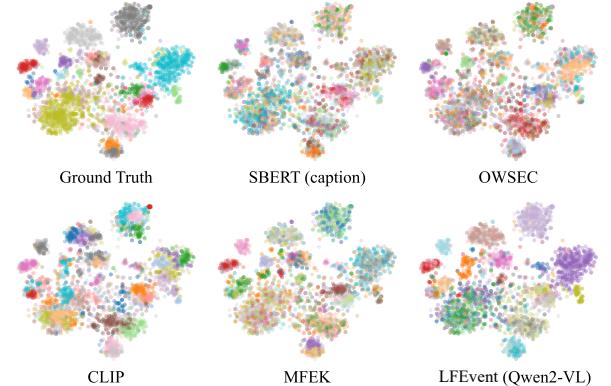
Overall, LFEEvent maintains performance above baseline methods across all tested settings, demonstrating robustness to hyperparameter variations and adaptability to diverse and noisy social media environments.

## 5.6 Visualization (Q5)

To qualitatively assess the clustering performance of LFEEvent, we visualize its results on the NED (S1) dataset using t-SNE, as shown in Figure 6. Each point represents a social media message, colored according to its ground-truth event label. The visualization compares LFEEvent (with Qwen2-VL for semantic enhancement) to several baselines: SBERT (caption), CLIP, OWSEC, and MFEK.

The t-SNE plots show that LFEEvent generates more compact clusters within event classes and better separation across classes than the baselines, aligning closely with the ground-truth distribution. This indicates LFEEvent's ability to capture nuanced event semantics in an unsupervised setting. In contrast, SBERT and CLIP exhibit overlapping clusters, suggesting weaker event differentiation. While OWSEC and MFEK show some improvement, they still struggle with fine-grained separation of novel events due to their reliance on supervised methods.

LFEEvent's label-free contrastive learning and structure entropy-guided clustering allow it to accurately group messages, even for unseen events, demonstrating its robustness in handling noisy, multimodal social media data. The visualizations are presented without



**Figure 6: (Q5) T-SNE visualization of clustering results on NED (S1). Each node denotes a message, and each colour denotes an event class.**

subplot borders for clarity, ensuring an intuitive comparison of clustering performance across methods.

## 6 CONCLUSION

In this work, we address the challenge of Multimodal Social Event Detection in open-world settings, where event types are unknown and social media data are noisy and diverse. We propose LFEEvent, the first label-free, clustering-based framework for MSED. LFEEvent uses self-supervised contrastive learning to jointly model text and images without predefined labels or manual annotations. A semantic enhancement mechanism constructs high-quality positive pairs, enabling robust modeling of fine-grained event semantics and adaptation to unseen events. Extensive experiments on multiple benchmarks show that LFEEvent consistently outperforms strong unimodal and multimodal baselines, effectively addressing two challenges: (1) learning discriminative multimodal representations from heterogeneous, noisy data, and (2) adapting to dynamically emerging events in open-world scenarios. However, LFEEvent shares common limitations of contrastive learning, such as sensitivity to hyperparameters and risk of representation collapse. Future work will extend LFEEvent to event evolution tracking and popularity prediction, advancing situational awareness and decision support in real-world social media environments.

## 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, the Fundamental Research Funds for the Central Universities.

## References

- [1] Mahdi Abavisan, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14679–14689.
- [2] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*.
- [3] Tilman Beck, Ji-Ung Lee, Christina Viehmann, Marcus Maurer, Oliver Quiring, and Iryna Gurevych. 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.
- [4] Yuwei Cao, Hao Peng, Jia Wu, Yingtong Dou, Jianxin Li, and Philip S Yu. 2021. Knowledge-preserving incremental social event detection via heterogeneous gnn. In *Proceedings of the Web Conference 2021*. 3383–3395.
- [5] Yuwei Cao, Hao Peng, Zhengtao Yu, and S Yu Philip. 2024. Hierarchical and incremental structural entropy minimization for unsupervised social event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8255–8264.
- [6] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15750–15758.
- [7] 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.
- [8] 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).
- [9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheisslghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.
- [10] Yuanyuan Guo, Zehua Zang, Hang Gao, Xiao Xu, Rui Wang, Lixiang Liu, and Jiangmeng Li. 2024. Unsupervised social event detection via hybrid graph contrastive learning and reinforced incremental clustering. *Knowledge-Based Systems* 284 (2024), 111225.
- [11] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 22105–22113.
- [12] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testugine. 2019. Supervised multimodal bitriformers for classifying images and text. *arXiv preprint arXiv:1909.02950* (2019).
- [13] Taiwo Kolajo, Olawande Daramola, and Ayodele A Adebiyi. 2022. Real-time event detection in social media streams through semantic analysis of noisy terms. *Journal of Big Data* 9, 1 (2022), 90.
- [14] Angsheng Li, Jiankou Li, and Yicheng Pan. 2015. Discovering natural communities in networks. *Physica A: Statistical Mechanics and its Applications* 436 (2015), 878–896.
- [15] Angsheng Li and Yicheng Pan. 2016. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory* 62, 6 (2016), 3290–3339.
- [16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [17] Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM on Web Conference 2024*. 2359–2370.
- [18] Zehang Lin, Jiayuan Xie, and Qing Li. 2024. Multi-modal news event detection with external knowledge. *Information Processing & Management* 61, 3 (2024), 103697.
- [19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
- [20] Frank Nielsen and Frank Nielsen. 2016. Hierarchical clustering. *Introduction to HPC with MPI for Data Science* (2016), 195–211.
- [21] Hao Peng, Jianxin Li, Qiran Gong, Yangqiu Song, Yuanxing Ning, Kunfeng Lai, and Philip S Yu. 2019. Fine-grained event categorization with heterogeneous graph convolutional networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 3238–3245.
- [22] Hao Peng, Ruitong Zhang, Shaoning Li, Yuwei Cao, Shirui Pan, and S Yu Philip. 2022. Reinforced, incremental and cross-lingual event detection from social messages. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 980–998.
- [23] Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2012. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. 1–8.
- [24] Daniela Pohl, Abdelhamid Bouchachia, and Hermann Hellwagner. 2015. Social media for crisis management: clustering approaches for sub-event detection. *Multimedia tools and applications* 74 (2015), 3901–3932.
- [25] Shengsheng Qian, Hong Chen, Dizhan Xue, Quan Fang, and Changsheng Xu. 2023. Open-world social event classification. In *Proceedings of the ACM Web Conference 2023*. 1562–1571.
- [26] Shengsheng Qian, Shengjie Zhang, Dizhan Xue, Huaiwen Zhang, and Changsheng Xu. 2025. Learning Temporal Event Knowledge for Continual Social Event Classification. *IEEE Transactions on Knowledge and Data Engineering* (2025).
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [28] Nils Reimers and Iryna Gurevych. 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.
- [29] Jiaqian Ren, Lei Jiang, Hao Peng, Yuwei Cao, Jia Wu, Philip S Yu, and Lifang He. 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.
- [30] Jiaqian Ren, Lei Jiang, Hao Peng, Zhiwei Liu, Jia Wu, and Philip S Yu. 2022. Evidential temporal-aware graph-based social event detection via dempster-shafer theory. In *2022 IEEE International Conference on Web Services (ICWS)*. IEEE, 331–336.
- [31] Jiaqian Ren, Hao Peng, Lei Jiang, Zhifeng Hao, Jia Wu, Shengxiang Gao, Zhengtao Yu, and Qiang Yang. 2024. Toward cross-lingual social event detection with hybrid knowledge distillation. *ACM Transactions on Knowledge Discovery from Data* 18, 9 (2024), 1–36.
- [32] Samir Elloumi Sihem Sahnoun and Sadok Ben Yahia. 2020. Event detection based on open information extraction and ontology. *Journal of Information and Telecommunication* 4, 3 (2020), 383–403. doi:10.1080/24751839.2020.1763007
- [33] Li Sun, Zhenhao Huang, Hao Peng, Yujie Wang, Chunyang Liu, and Philip S Yu. 2024. LSEnet: Lorentz Structural Entropy Neural Network for Deep Graph Clustering. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.
- [34] Zhongqing Wang and Yue Zhang. 2017. A Neural Model for Joint Event Detection and Summarization. In *IJCAI*. 4158–4164.
- [35] Feng Xue, Richang Hong, Xiangnan He, Jianwei Wang, Shengsheng Qian, and Changsheng Xu. 2019. Knowledge-based topic model for multi-modal social event analysis. *IEEE Transactions on Multimedia* 22, 8 (2019), 2098–2110.
- [36] Zhenguo Yang, Qing Li, Zheng Lu, Yun Ma, Zhiguo Gong, and Haiwei Pan. 2015. Semi-supervised multimodal clustering algorithm integrating label signals for social event detection. In *2015 IEEE International Conference on Multimedia Big Data*. IEEE, 32–39.
- [37] Zhiwei Yang, Yuecen Wei, Haoran Li, Qian Li, Lei Jiang, Li Sun, Xiaoyan Yu, Chunming Hu, and Hao Peng. 2024. Adaptive differentially private structural entropy minimization for unsupervised social event detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2950–2960.
- [38] Chen Yu, Bin Hu, and Zhiguo Wang. 2025. Open-world disaster information identification from multimodal social media. *Complex & Intelligent Systems* 11, 1 (2025), 7.
- [39] Xiaoyan Yu, Jiaqian Ren, Lei Jiang, Hao Peng, Zhifeng Hao, Li Sun, Kun Peng, Liehuang Zhu, and Philip S Yu. 2025. PromptSED: An evolving topic-enhanced prompting framework for incremental social event detection. *Neural Networks* (2025), 107772.
- [40] Yongshe Yu, Jia Wu, and Jian Yang. 2023. Social Event Detection with Reinforced Deep Heterogeneous Graph Attention Network. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 463–472.
- [41] Jingyun Zhang, Hao Peng, Li Sun, Guanlin Wu, Chunyang Liu, and Zhengtao Yu. 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.
- [42] Kun Zhang, Xiaoyan Yu, Pu Li, Hao Peng, and Philip S Yu. 2024. SocialED: A Python Library for Social Event Detection. *arXiv preprint arXiv:2412.13472* (2024).
- [43] Sicheng Zhao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. 2017. Real-time multimedia social event detection in microblog. *IEEE transactions on cybernetics* 48, 11 (2017), 3218–3231.
- [44] Han Zhou, Hongpeng Yin, Hengyi Zheng, and Yanxia Li. 2020. A survey on multi-modal social event detection. *Knowledge-Based Systems* 195 (2020), 105695.