Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061037

ARTICLE

Event-Driven Attention Network: A Cross-Modal Framework for Efficient Image-Text Retrieval in Mass Gathering Events

Kamil Yasen[1,#], Heyan Jin[2,#], Sijie Yang[2], Li Zhan[2], Xuyang Zhang[2], Ke Qin[1,3] and Ye Li[2,3,*]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China
[2]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China
[3]Kashi Institute of Electronics and Information Industry, Kashi, 844508, China

*Corresponding Author: Ye Li. Email: liyeuestc@uestc.edu.cn

[#]Both Kamil Yasen and Heyan Jin contributed equally to this work

Received: 15 November 2024; Accepted: 09 January 2025; Published: 16 April 2025

ABSTRACT: Research on mass gathering events is critical for ensuring public security and maintaining social order. However, most of the existing works focus on crowd behavior analysis areas such as anomaly detection and crowd counting, and there is a relative lack of research on mass gathering behaviors. We believe real-time detection and monitoring of mass gathering behaviors are essential for migrating potential security risks and emergencies. Therefore, it is imperative to develop a method capable of accurately identifying and localizing mass gatherings before disasters occur, enabling prompt and effective responses. To address this problem, we propose an innovative Event-Driven Attention Network (EDAN), which achieves image-text matching in the scenario of mass gathering events with good results for the first time. Traditional image-text retrieval methods based on global alignment are difficult to capture the local details within complex scenes, limiting retrieval accuracy. While local alignment-based methods are more effective at extracting detailed features, they frequently process raw textual features directly, which often contain ambiguities and redundant information that can diminish retrieval efficiency and degrade model performance. To overcome these challenges, EDAN introduces an Event-Driven Attention Module that adaptively focuses attention on image regions or textual words relevant to the event type. By calculating the semantic distance between event labels and textual content, this module effectively significantly reduces computational complexity and enhances retrieval efficiency. To validate the effectiveness of EDAN, we construct a dedicated multimodal dataset tailored for the analysis of mass gathering events, providing a reliable foundation for subsequent studies. We conduct comparative experiments with other methods on our dataset, the experimental results demonstrate the effectiveness of EDAN. In the image-to-text retrieval task, EDAN achieved the best performance on the R@5 metric, while in the text-to-image retrieval task, it showed superior results on both R@10 and R@5 metrics. Additionally, EDAN excelled in the overall Rsum metric, achieving the best performance. Finally, ablation studies further verified the effectiveness of event-driven attention module.

KEYWORDS: Mass gathering events; image-text retrieval; attention mechanism

1 Introduction

Mass gathering events are defined as large-scale gatherings of people for a common purpose, including political rallies, protests, concerts, festivals, and emergency evacuations. The phenomenon of such gatherings has garnered considerable attention from researchers and policymakers in recent times. Investigating mass gatherings is crucial for understanding contemporary social dynamics and enhancing public policy [1].

Since the advent and widespread adoption of deep learning algorithms, research on mass gathering events has employed methods such as Convolutional Neural Networks (CNNs) [2,3] and Long Short-Term Memory (LSTM) networks [4] to identify crowd behavior. Examples include crowd counting [5–8], crowd anomaly detection [9–11], and pedestrians tracking [12–15]. Crowd anomaly detection focuses on identifying abnormal behaviors such as fighting [16], robbery [17], and sudden running [9]. However, the existing research predominantly emphasizes detecting the current state of crowds while overlooking a more profound understanding of the mass gathering events that precipitate these states.

Mass gathering events often precede crowd-related disasters or incidents, representing a significant anomaly [18]. Consequently, timely detection and real-time monitoring of crowd behavior are essential for preventing such occurrences. In this paper, we explore research on mass gathering events. Our objective is to design a method that can accurately identify the spatial location of large-scale gatherings before significant crowd disasters occur, thus enabling real-time monitoring and proactive prevention.

Image-text retrieval, a method that integrates visual and textual features, enables the retrieval of relevant images based on image descriptions or or the extraction of related descriptions from images. This method has emerged as an effective way to our goal. The core of image-text retrieval technology lies in aligning images with text descriptions to uncover semantic similarities and establish meaningful connections between them. To our knowledge, this study represents the first application of image-text retrieval in the domain of mass gathering events. Through image-text retrieval technology, we can efficiently match and analyze images and texts related to mass gathering events, rapidly acquiring multimodal information relevant to these incidents. For instance, in practical applications, alerts about mass gathering events allow for the swift retrieval of corresponding surveillance footage using textual scene descriptions. This enables accurate identification of the gathering's location, facilitating rapid response measures and enhancing overall operational efficiency. Additionally, by pre-setting textual descriptions of various gathering scenarios to create a database, we can match live video frames from cameras with the texts stored in this database. A successful match indicates the occurrence of a gathering, allowing for immediate attention. These detailed descriptions of gathering scenes not only aid personnel in conducting thorough analyses but also in predicting crowd movements, offering vital insights for predicting the event's trajectory.

The development of image-text retrieval algorithms has evolved from traditional feature matching techniques to advanced deep learning approaches. Currently, mainstream deep learning methods can be broadly categorized into two types: global alignment methods and local alignment methods. Early global alignment methods [19,20] involved mapping different modalities into a joint embedding space by designing cross-modal alignment loss functions (Fig. 1a). However, these global alignment methods may overlook the local information inherent in both images and text, as well as the interactions between modalities, thereby limiting the model's capabilities. In response to these limitations, subsequent methodologies [21,22] have enhanced the cross-modal interaction abilities and improved retrieval accuracy by establishing corresponding relationships between image regions and textual words. These advancements are collectively referred to as local alignment methods (Fig. 1b). While local alignment approaches have matured significantly, they often neglect the relationships between semantic objects due to the complexities associated with visual semantic differences. Therefore, recent research efforts have concentrated on employing Transformer models [23] or graph convolutional networks [24] to effectively model relationships among semantic objects within multi-modal data framework. This focus aims to facilitate more efficient image-text retrieval processes.

However, the application of the aforementioned methods to research on mass gathering events continues to pose numerous challenges. Firstly, the image and text data associated with mass gathering events are highly diverse and complex. Relying solely on global alignment or local alignment methods is insufficient for fully capturing the information embedded within these image or text datasets. Moreover, existing local

alignment methods that model multi-modal semantic object relationships typically employ transformers to facilitate interactions between each image region feature and each word-level text feature. This approach not only diminishes retrieval efficiency but also fails to adequately account for the unique characteristics of mass gathering events as well as the correspondence among images, texts, and event types. Consequently, there is an urgent need to develop an image-text retrieval method tailored to the unique characteristics of mass gathering events.

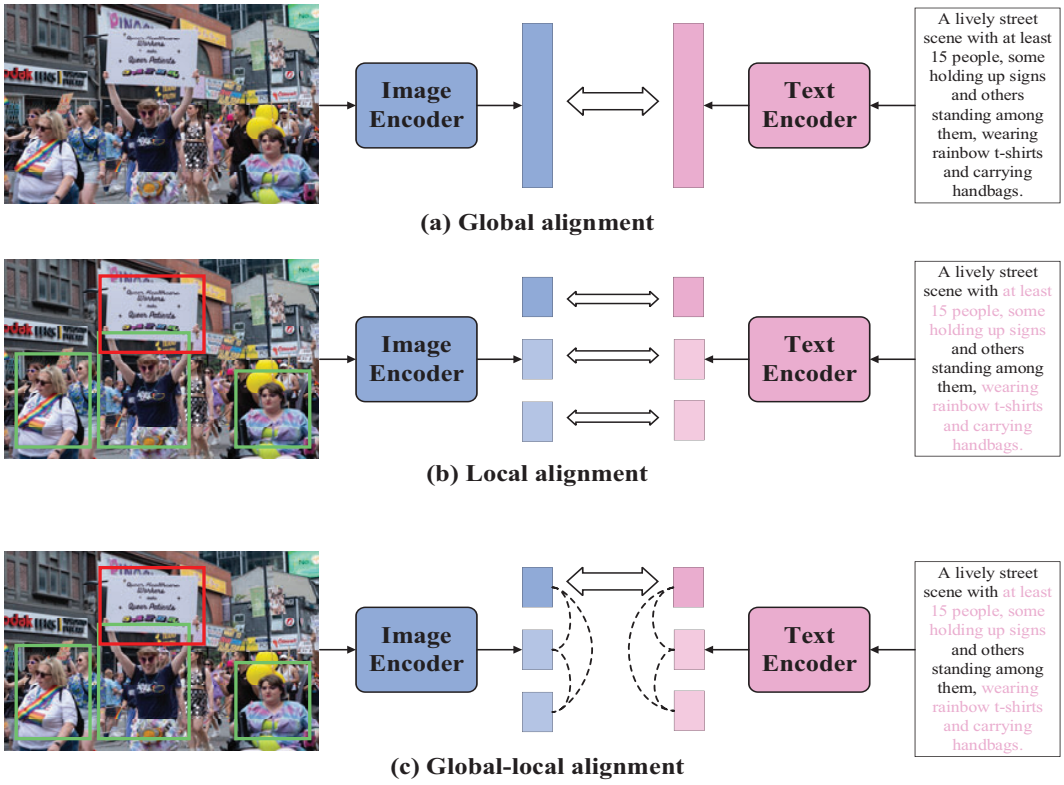

Figure 1: Illustration of different feature alignment methods

To address the aforementioned issues, this paper first integrates both global and local alignment methods (Fig. 1c), enabling the model to align the global semantic representations of images and texts while also capturing more detailed local features, thereby achieving enhanced modal interactions. Subsequently, We propose an Event-Driven Attention Network (EDAN) to model the relationships between semantic objects. This module utilizes event types as guiding labels, directing the alignment of different modal features based on these event types during both global and local alignment stages. To validate the effectiveness of our proposed method, we specifically constructed a dedicated multi-modal dataset for studying mass gathering events. This dataset comprises a substantial number of images related to mass gatherings, each accompanied by five descriptive captions. We applied our method to this dataset and achieved satisfactory results. This work pioneers the application of image-text retrieval methods to the study of mass gathering events, laying a solid foundation for future in-depth analyses and processing of such events. The main contributions can be summarized as follows:

(1) This study pioneers the application of image-text retrieval to the study of mass gathering events, bridging a notable gap in multimodal research within this domain. We also develop a specialized dataset to both validate our approach and support future research endeavors.

(2) We enhance the precision of image-text correspondence by integrating global with local alignment strategies. This amalgamation of global perspectives and local details during the matching process enables a more holistic capture of the semantic relationships between the two modalities.

(3) We propose an innovative event-driven attention network, specifically designed for the unique characteristics of mass gathering events. This network steers the alignment of local features, allowing the model to focus more effectively on words or regions that are relevant to the type of event, thereby improving the efficiency of retrieval.

2 Related Work

2.1 Crowd Behavior Analysis

Crowd is a group of people gathered in a certain location. It represents a unique assembly of individuals who share a common physical space. With the continuous growth of the global population, the frequency and scale of crowds in public areas have significantly increased. This trend underscores the importance of analyzing crowd behavior in such environments. Video surveillance has proven to be an effective method for studying crowd dynamics in crowded scenes. Key tasks in crowd analysis include understanding crowd behavior, tracking movement patterns, estimating crowd density, and detecting unusual motion.

In recent years, deep learning has made significant strides across various fields, including the understanding of crowd behavior. Feng et al. [25] employed PCANet [26] to extract event features and utilized a Deep Gaussian Mixture Model (Deep GMM) to identify abnormal events. Yang et al. [27] proposed a novel deep learning architecture, DeepSDAE, for anomaly detection, which can be trained using reinforcement learning. Zhang et al. [28] developed an innovative LSTM-based framework to model interactions between pedestrians and their environment, enabling the prediction of both individual and group behavior trajectories. Su et al. [29] detected social groups based on interpersonal distances and spatiotemporal trajectories of pedestrians. Alafif et al. [30] introduced a GAN-based approach for detecting anomalous behavior at the individual level. In their work, the GAN was trained exclusively on normal samples, where the generator simulates normal inputs and the discriminator distinguishes between real and generated samples. As a result, the discriminator can effectively identify anomalies when presented with abnormal inputs.

2.2 Image-Text Retrieval

Image-Text Retrieval refers to the similarity search between two modalities (text and image) to establish the connection and achieve interoperation between the two types of data. This field encompasses two main tasks: image-to-text retrieval and text-to-image retrieval. The image-to-text retrieval involves finding matching textual descriptions based on a given image, whereas retrieving images from text involves finding matching images based on a given text description.

As an emerging field, image-text retrieval has considerable research value. Especially, with the rapid development of natural language and computer vision, image-text retrieval has obtained rapid development [31]. By summarizing the image-text retrieval method in recent years, they can be roughly divided into two categories: global alignment method and local alignment method.

Global alignment methods. Global alignment methods focus on mapping both image and text data into a shared feature space, enabling direct comparisons and retrievals based on overall content similarity. These methods aim to capture the semantic correspondence between the visual and textual modalities by

projecting them into a common embedding space. Once both images and texts are represented within this shared space, their similarity can be directly measured, which facilitates efficient retrieval of semantically related pairs. A notable example of such techniques is the work by Faghri et al. [20], who introduced the use of contrastive loss to enhance the learning of a shared embedding space for images and texts. In their approach, the contrastive loss encourages matching image-text pairs to be closer together in the shared space, while pushing mismatched pairs farther apart. This method has been influential in advancing the field of cross-modal retrieval by improving the discriminative power of the learned representations. Recent studies, Meng et al. [19] proposed a Prototype Local-Global Alignment Network, which goes beyond global alignment by incorporating both fine-grained (local) and holistic (global) levels of matching. Their method ensures comprehensive cross-modal alignment by simultaneously considering detailed feature correspondences at the local level, such as object and word interactions, alongside broader, global content alignment. This dual-level alignment strategy enhances the model's ability to capture intricate relationships between image regions and textual components, leading to more accurate and robust cross-modal retrieval performance.

Local alignment methods. Local alignment methods in image-text retrieval aim to align specific regions within images with corresponding segments of text, providing a more granular and detailed matching process. Unlike global alignment, which focuses on holistic similarity, local alignment captures the finer details and relationships between visual objects in the image and their corresponding textual descriptions. This approach ensures that individual components, such as specific objects, actions, or attributes in the image, are accurately matched with relevant words or phrases in the text, leading to more precise retrieval. Early works like Karpathy et al. [21] utilized DNNs to align image regions with corresponding textual descriptions. Their method breaks down the image into a set of regions and the text into a sequence of words, then uses a neural network to identify the most relevant pairs of regions and words. Anderson et al. [32] proposed a bottom-up and top-down attention mechanism that aligns objects detected in images with text segments. In their framework, the bottom-up attention identifies objects in the image using object detection methods, while the top-down attention focuses on aligning these detected objects with specific text segments based on their importance and relevance. Lee et al. [22] use different important image regions and words as contexts to infer the similarity between image texts, namely the stacked cross attention mechanism: for a given image and text, the words in the sentences related to each image region are first processed, and the information of the words appearing in each image region and sentence is compared to determine the importance of the image region. In recent years, there has been a growing focus on aligning prototype representations at both local and global levels. For example, the method described in [33] incorporates a Semantic Alignment Module (SAM), which captures and aligns semantic information between the visual and textual domains. SAM ensures that both fine-grained (local) details and higher-level (global) representations are aligned, allowing the model to capture a more comprehensive semantic relationship between images and texts.

2.3 Image-Text Retrieval Datasets

Image-text retrieval relies on diverse datasets to evaluate the effectiveness of various models. With the rapid development of information technology today, various sources of information have become widespread, supporting the development of image and text retrieval datasets. The Flickr30k [34] dataset consists of 31,783 images collected from Flickr, each paired with five textual descriptions. This dataset is widely used for benchmarking image captioning and retrieval models. The MSCOCO [35] dataset contains over 330,000 images, each annotated with five textual descriptions. Google Conceptual Captions (GCC) [36] is a large dataset with 3.3 million image-caption pairs extracted from the web. The captions are generated from the alt-text of images, making this dataset diverse and extensive for training and evaluating models.

These datasets provide extensive and diverse data for training and evaluating image-text retrieval models, contributing significantly to advancements in the field.

3 Method

In this section, we provide a detailed overview of EDAN, Fig. 2 illustrates the overall architecture. We begin by introducing the feature extraction stage for image and text in Section 3.1. In Section 3.2, we explain the contextual and global representation learning process for image. In Section 3.3, we describe the process of event-driven semantic enhancement, with a detailed introduction to our proposed event-driven attention module. Section 3.4 outlines the cross-modal feature alignment stage. Finally, in Section 3.5, we introduce the objective function of the overall framework and the inference process.

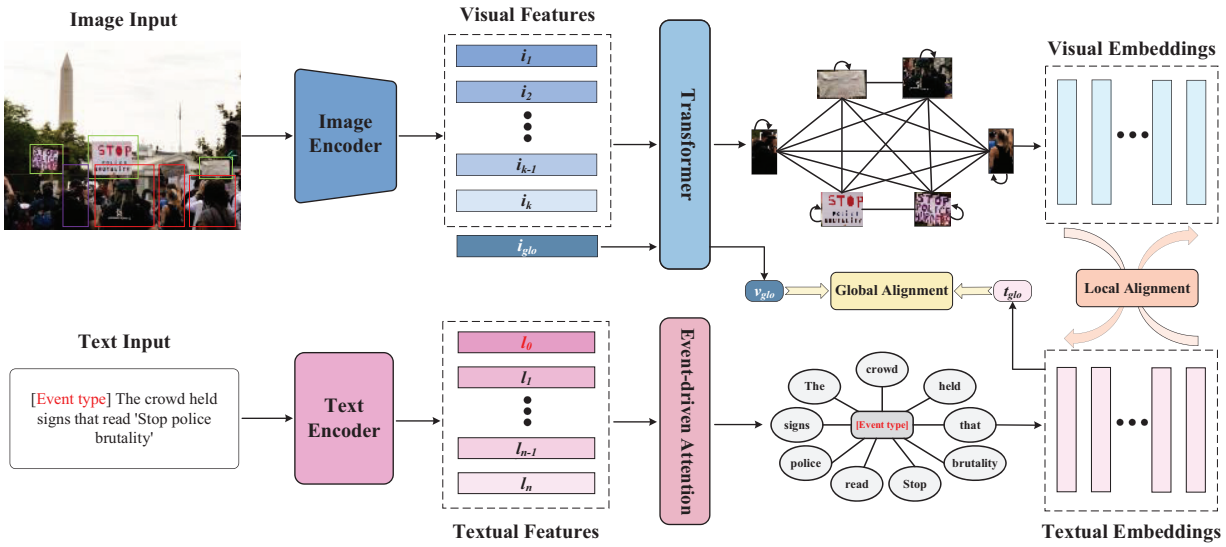

Figure 2: The image encoder initially employs Faster R-CNN to extract features from salient regions within the image, as well as global feature that represent the entire image. These extracted features are subsequently processed through a transformer encoder to learn contextual and global representations, resulting in contextual embeddings for each region alongside a comprehensive global representation. For the text component, both the textual data and event label are first transformed into vector sequences using a text encoder (Word2Vec), which generates a sequence of word embeddings along with an event vector. These embeddings are then input into an encoder equipped with an event-driven attention module designed for event-guided semantic enhancement, ultimately yielding a semantically enriched word-level feature sequence. During the global alignment process, the word-level feature sequence is aggregated through pooling techniques to derive the overall textual feature representation. This representation is directly compared with the global representation of the image to assess similarity. In terms of local alignment, each region-specific feature of the image interacts directly with the word-level features of the text to compute their similarity

3.1 Multi-Modal Feature Extraction

To extract features from both image and text modalities, we employ separate pre-trained encoders tailored for each modality.

Image Encoder. For the image data, we employ Faster R-CNN [37], a region-based convolutional neural network which integrates bottom-up and top-down attention (BUTD) [32] to detect salient regions within images. This model is pre-trained on the Visual Genome [38]. By leveraging this pre-trained architecture, our system can effectively capture essential visual details and structural information from the images, thereby enabling it to address the complexities of visual content associated with mass gathering events.

As shown in Fig. 2, for an input image I, the Faster R-CNN will extract region-level features $r_i (i = 1, 2, \ldots, K)$, where K represents the number of obvious regions of the image. These features are then processed through a fully connected layer to convert them into d-dimensional vectors, resulting in the visual feature sequence $\{i_1, i_2, \ldots, i_K\}$, $i_j \in R^d$, $(j = 1, 2, \ldots, K)$. Finally, these features are combined with the global feature vector i_0 to obtain the final visual feature sequence $I = \{i_0, i_1, i_2, \ldots, i_K\}$.

Text Encoder. In recent studies, BERT [39] (Bidirectional Encoder Representations from Transformers) has gained widespread adoption for the extraction of textual features. However, BERT primarily emphasizes modeling the relationships between words and their surrounding context within a sentence, as well as the connections between individual words and the overall text context. This focus tends to overlook, to some extent, the semantic relationships among words. We contend that in the task of image-text matching, comprehending the semantics of text is equally vital, as it directly influences the effectiveness of subsequent feature alignment.

Word2Vec [40], a well-established model for learning word embeddings, is recognized for its capability to capture both semantic and syntactic relationships among words by utilizing local context within a sliding window framework. Consequently, in this paper, we employ Word2Vec to extract textual features at a semantic level.

Specifically, for the input text T, we split it into a set of words $T = \{w_1, w_2, \ldots, w_N\}$, where N represents the number of words contained in the text. Each word in the text is then processed by a pre-trained Word2Vec model to obtain its corresponding word embedding $H = \{h_1, h_2, \ldots, h_N\}$, where $h_i, (i = 1, 2, \ldots, N)$ is the embedding vector for the i-th word. Finally, a fully connected layer performs a linear transformation on H:$L = W \cdot H + b$, $L = \{l_1, l_2, \ldots, l_N\}$, where W represents the weight matrix and b denotes the bias vector, $l_i \in \mathrm{R}^{\mathrm{d}}$, $l_i \in \mathrm{R}^{\mathrm{d}}$, $(i = 1, 2, \ldots, N)$. The resulting feature sequence serves as a representation of the textual features.

3.2 Visual Contextual and Global Representation Learning

Mass gathering events frequently involve dynamic scenes and complex interactions among multiple elements, placing significant demands on image-text retrieval models, particularly in terms of fine-grained retrieval capabilities. In such contexts, the retrieval model must go beyond identify individual components within images to develop a deep understanding of the intricate relationships between these components and the broader event context. Contextual representation learning plays a crucial role in this process by seamlessly integrating visual features from images with semantic information derived from text. This enables the model to capture subtle cross-modal semantic relationships and contextual dependencies more effectively. Such advanced information integration markedly enhances the accuracy of image-text matching and significantly improves the overall performance of the retrieval system in practical applications. Moreover, contextual representation learning equips the model to handle scenarios characterized by dynamic changes and rich contextual information with greater adaptability and predictive power. This capability is essential for processing real-time information flows and delivering highly relevant search results in dynamic, real-world environments.

In recent years, the attention-based architectures have demonstrated outstanding performance in both vision and language tasks. The attention mechanism plays a critical role in contextual representation learning by focusing on the most relevant information within input sequence to enhance the quality of representations. Specifically, it enables the model to dynamically assign weights to different regions or words, ensuring that more pertinent contextual information is highlighted and facilitating the capture of complex, long-range dependencies. In this work, we employ the classic transformer encoder [41] architecture as the visual contextual encoder for embedding image region features.

In addition, most fine-grained models leverage cross-attention mechanisms to dynamically align each element with corresponding elements from another modality. While these models generally achieve superior performance compared to their coarse-grained counterparts, the extensive cross-modal computations required between images and text significantly reduce computational efficiency. This inefficiency poses challenges to scalability and flexibility, both of which are critical for practical applications. To address this issue, an efficient method is needed to maintain fine-grained retrieval while enhancing computational efficiency. Motivated by this, we propose a learnable visual global representation that can encapsulate the entire image. During the subsequent global alignment phase, this representation enables direct similarity computation with the global textual representation, thereby significantly improving retrieval efficiency.

To get the visual contextual embeddings, we use the visual contextual encoder, which is composed of six layers of standard transformer encoder. Each layer of the encoder consists of a multi-head self-attention mechanism and a feed-forward layer. As shown in Fig. 2, for the visual feature sequence $I = \{i_0, i_1, i_2, \ldots, i_K\}$ extracted during the feature extraction phase, we use the sequence as input to the visual contextual encoder. The input sequence incorporates contextual information to produce the visual embedding sequence $\widehat{V} = \{\widehat{v}_0, \widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_K\}$, where K represents the sequence length:

$$\widehat{V} = Transformer(I) = \{\widehat{v}_0, \widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_K\} \tag{1}$$

where $\widehat{v}_0$ is the learned global image embedding, and $\widehat{v}_1, \ldots, \widehat{v}_K$ are the local representations about regions.

3.3 Event-Driven Semantic Enhancement

In fine-grained retrieval, the descriptive text for images often contains ambiguity and redundant information, which complicates the accurate extraction of core semantics relevant to the image content using raw text features. Furthermore, during the subsequent feature alignment stage, each word-level feature must participate in cross-attention operations. Directly applying these operations to raw text features can significantly increase computational overhead, thereby reducing both efficiency and precision. By leveraging semantic enhancement methods, it becomes possible to more effectively extract key semantic information that aligns with the target image while filtering out irrelevant or noisy features. This approach significantly improves the accuracy and performance of cross-modal alignment and retrieval.

Additionally, our investigation into mass gathering events reveals that the variety of event types is relatively limited, primarily encompassing protest activities, festival celebrations, religious gatherings and similar occurrences. Consequently, we propose a textual semantic enhancement method tailored to the characteristic of limited types of mass gathering events: Event-driven Attention. Specifically, we treat these event types as guiding labels that enable the model to focus on concepts pertinent to their corresponding event type. For instance, when analyzing protest activities, the model should emphasize areas depicting expressions of anger within crowds or relevant slogans; conversely, during celebratory events such as festivals or parades, emphasis might be placed on scenes featuring fireworks or lanterns.

Event-Driven Attention. As shown in Fig. 2, for an image-text pair, we first create an event label corresponding to its event type and embed it into the text input. The details of the event label will be explained in detail in Section 4. Upon processing through Word2Vec, this label yields a corresponding feature which will be referred to as event feature l_0. This event feature is then integrated with other word-level features to create a comprehensive feature sequence $\{l_0, l_1, l_2, \ldots, l_N\}$. The feature sequence is then as input into the event-driven attention module, as illustrated in Fig. 3. We separate the event feature l_0 from other features. We input l_0 into layer normalization to obtain the query Q. Meanwhile, all other features undergo layer

normalization to produce the key K, and value V

$$Q = l_0 W^Q,\ K = LW^K,\ V = LW^V \tag{2}$$

where W^Q, W^K, $W^V \in R^{d\times d}$ are the trainable parameters. Then, the key vector K only needs to perform a dot product with the query vector Q to get the attention scores, which consist of a single event feature. This approach significantly reduces computational load. Next, these scores are processed through the softmax function to obtain the attention weights

$$Attention\_weights = softmax(\frac{QK^T}{\sqrt{d}}) \tag{3}$$

where d is the dimension of the key vectors and is used for scaling to prevent the dot product from becoming too large. The attention weights are then used to weight the value vectors

$$\widehat{F} = softmax(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

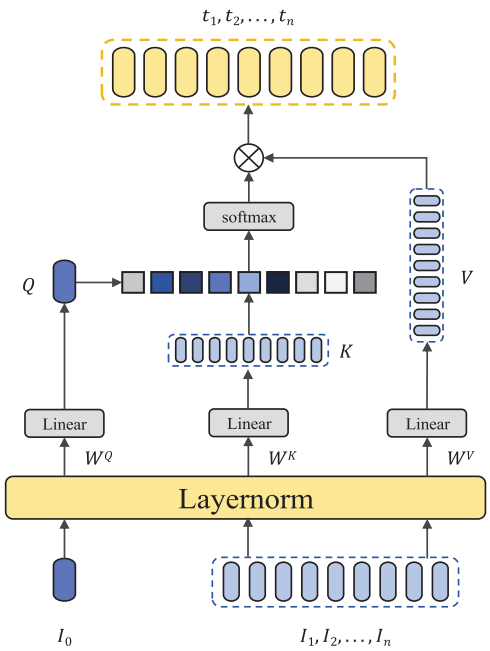

Figure 3: Illustration of the Event-Driven Attention (EDA) module. The EDA only computes the query vector for the event label, significantly reducing the computational load

Ultimately resulting in the textual embedding sequence $T = \{t_1, t_2, \ldots, t_N\}$, where N represents the length of the text.

3.4 Cross-Modal Feature Alignment

After obtaining the contextual embeddings of image and the enhanced semantic embeddings of text, an alignment strategy is essential for fully integrating these modalities to achieve cross-modal understanding and matching. However, in the context of mass gathering events, global alignment—while effective in

capturing the overall trends and emotions associated with such events—often neglects the richness of local details. For instance, during a protest event, global alignment may identify a prevailing sentiment of anger but might overlook specific expressions from individual participants or significant slogans that are crucial for comprehending the nature of the event. Conversely, while local alignment emphasizes detailed features, it lacks an appreciation for the overarching structure, which can result in a fragmented interpretation of the event. For example, in a celebratory gathering, focusing solely on localized elements such as fireworks or decorations may fail to capture interactions among participants or shifts in collective emotion. Therefore, relying exclusively on either global or local alignment results in information loss that limits comprehensive analysis and understanding of mass gathering events. By integrating both alignment strategies, we can more effectively capture the complexity and diversity inherent to such events, thereby enhancing our model's capability to recognize and analyze various types of gatherings.

Global alignment. During the global alignment phase, it is necessary to compute similarity using the global feature of both the image I and text T. For image, the feature learned during the visual global representation learning stage can be directly used as global feature. For text, however, a pooling operation is required on the enhanced semantic feature sequence to obtain the global representation. In this study, we use average pooling

$$v_{glo} = \widehat{v}_0,\ t_{glo} = \frac{1}{N}\sum_{j=1}^{N} t_j \tag{5}$$

So the semantic similarity $\mathrm{sim}(I,\ T)$ between the entire image and text can be calculated by:

$$sim(I,\ T) = \frac{W_s^g |v_{glo} - t_{glo}|^2}{\parallel W_s^g |v_{glo} - t_{glo}|^2 \parallel_2} \tag{6}$$

Given a batch of B image-text pairs, we can obtain the global feature pairs $\{(v_{glo},\ t_{glo})\}_{i=1}^{B}$ for both the images and texts. The probability of a matching pair can be computed using the following formula:

$$p_{i,j}(I,\ T) = \frac{\exp(sim(v_{glo,i},\ t_{glo,j})/\tau)}{\sum_{k=1}^{B} exp(sim(v_{glo,i},\ t_{glo,k})/\tau)},\ p_{i,j}(T,\ I) = \frac{\exp(sim(t_{glo,j},\ v_{glo,i})/\tau)}{\sum_{k=1}^{B} exp(sim(t_{glo,k},\ v_{glo,i},)/\tau)} \tag{7}$$

where τ is a learnable temperature parameter which adjusts the smoothness of the output probability distribution. Subsequently, we compute the contrastive loss for both the image and text:

$$\mathcal{L}_{i2t,glo} = \frac{1}{B}\sum_{i=1}^{B} H\left(y_{i,j},\ p_{i,j}\left(I,T\right)\right),\ \mathcal{L}_{t2i,glo} = \frac{1}{B}\sum_{i=1}^{B} H\left(y_{i,j},\ p_{i,j}\left(T,\ I\right)\right) \tag{8}$$

where $H(\cdot,\cdot)$ is the cross-entropy function, and y is the matching label, where $y_{i,j} = 1$ indicates that the image and text form a matching pair, while $y_{i,j} = 0$ indicates a non-matching pair. Then, we can compute the ITC loss function of global alignment:

$$\mathcal{L}_{itc,glo} = \left(\mathcal{L}_{i2t,glo} + \mathcal{L}_{t2i,glo}\right)/2 \tag{9}$$

Local alignment. At this stage, our focus is on the interaction between visual region features and textual word-level features. Specifically, in the task of calculating similarity from text T to image I for a given image-text pair. For textual feature sequence $T = \{t_1,\ t_2, \ldots,\ t_N\}$ and visual feature sequence $\widehat{V} = \{\widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_K\}$, we first compute the similarity $s(t_i,\ I)$ between each word-level feature t_i and the image I:

$$s\left(t_i,\ I\right) = S(t_i,\ \widehat{t}_i) \tag{10}$$

where $\widehat{t_i} = \sum_{j=1}^{K} \omega_{ij}\widehat{v}_j$, ω_{ij} is the weighting factor of $\widehat{v}_j$. By defining $s_{ij} = \frac{t_i^T \cdot \widehat{v}_j}{\|t_i\| \cdot \|\widehat{v}_j\|}$ as the similarity between the t_i and the $\widehat{v}_j$, ω_{ij} is typically positively associated with s_{ij}. S is the similarity calculation function, in our study, we use the cosine similarity function:

$$S(t_i, \widehat{t_i}) = \frac{t_i^T \cdot \widehat{t_i}}{\| t_i \| \cdot \| \widehat{t_i} \|} \tag{11}$$

Finally, we pool the similarities of all words to get the similarity score between the text T and the image I:

$$S(T, I) = \frac{1}{N} log \sum_{i=1}^{N} exp(NS(t_i, I)) \tag{12}$$

Similarly, the similarity from image to text can be obtained as:

$$S(I, T) = \frac{1}{K} log \sum_{j=1}^{K} exp(KS(\widehat{v}_j, T)) \tag{13}$$

When calculating the loss during the local alignment phase, we utilize the online hard negative mining hinge-based bidirectional triplet ranking loss, a technique also employed in VSE++ [20] and [42]. The fundamental principle underlying this loss function is to ensure that words within a text description are closely associated with their most relevant image regions when the image and sentence are correctly matched. Simultaneously, it guarantees that words and their corresponding regions in mismatched image-sentence pairs remain sufficiently distant in the feature space.

The online hard negative mining component further enhances the effectiveness of the loss function. During training, rather than treating all negative examples equally, the model focuses on the most challenging "hard negatives"—those mismatched image-text pairs that are still relatively similar and therefore more likely to be confused. By imposing a greater penalty for these hard negatives, the model learns to distinguish between very similar yet incorrect pairs, leading to improved generalization and enhanced retrieval accuracy. Mathematically, the local alignment loss function is defined as follows:

$$\mathcal{L}_{loc} = \sum_{(T,I)} [\alpha + S(T, \widehat{I}) - S(T, I)]_+ + [\alpha + S(\widehat{T}, I) - S(T, I)]_+ \tag{14}$$

Here, α is the margin parameter, $[x]\_\{+\} \equiv \max(x, 0)$. $\widehat{I} = \arg\max_{I' \neq I} s(T, \widehat{I})$ and $\widehat{T} = \arg\max_{T' \neq T} s(\widehat{T}, I)$ denote the most difficult image and sentence to distinguish within a training mini-batch, respectively.

3.5 Objective Function and Inference

In the training phase, EDAN integrates both global alignment and local alignment. This dual approach not only enhances the overall matching accuracy between images and texts but also improves the fine-grained matching of local regions and word-level features. The comprehensive objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{loc} + \mathcal{L}_{itc,glo} \tag{15}$$

Inference. During the inference process, the calculation of image-text similarity can be conducted using either the similarity from the global alignment phase or the similarity from the local alignment phase. Global alignment generally provides faster inference speeds but with relatively lower retrieval accuracy, whereas

local alignment typically achieves higher retrieval accuracy at the expense of slower inference times. In practical applications, such as mass gathering events, both high processing speed and high accuracy are essential. Therefore, we adopt the method proposed in [42], which combines the global alignment similarity and the local alignment similarity to obtain the final similarity score:

$$S_{infer}(I,\ T) = (1-\theta)sim(I,\ T) + \theta S(I,\ T), S_{infer}(T,\ I) = (1-\theta)sim(T,\ I) + \theta S(T,\ I) \tag{16}$$

4 Construction of Dataset

Most existing image-text retrieval datasets are constructed by using online resources. Although these datasets boast a large volume, their data sources span various domains, which complicates their application in specific fields. To address this issue, we construct a dedicated cross-modal retrieval dataset (MG) tailored for mass gathering events, which is created through four steps: (1) filtering part of the existing image-text retrieval datasets, (2) constructing part of the dataset from web resources, (3) expanding and merging; and (4) event label creation.

4.1 Filtering Part

Although existing datasets cannot be directly used to study mass gathering events, many of the images and descriptions they contain align with the characteristics of these events. Consequently, these resources can be filtered for effective utilization. In this section, we will outline the process of filtering the existing dataset. The overall process is shown in Fig. 4, which is divided into three steps: (1) dictionary construction, (2) text filtering, and (3) obtaining corresponding images based on the filtered text.

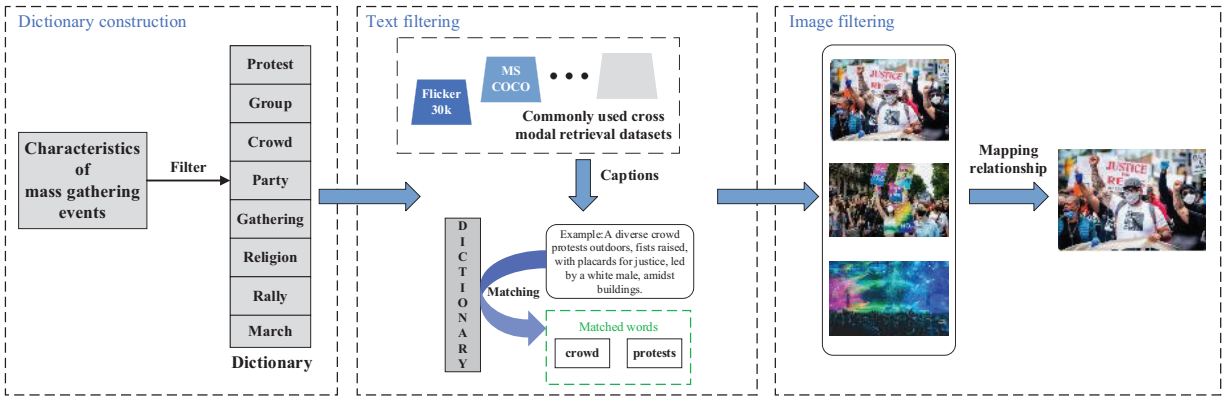

Figure 4: Production process of Part 1 (Filtering Part) of MG

Dictionary Construction. Based on the characteristics of mass gathering events, we have organized several keywords—such as group, gathering, protest, crowd, and violence—into a dictionary specifically for these events.

Text Filtering. We first match the image descriptions from general image-text retrieval datasets (such as Flicker30k [34], MSCOCO [35], etc.) with the dictionary constructed in Dictionary Construction. The matched texts serve as a preliminary filtered set, which is subsequently subjected to manual screening to identify those that exhibit characteristics indicative of mass gathering events. This process yields the final results for text filtering.

Image Filtering. Based on the texts filtered in Text Filtering and the established mapping relationship between texts and images, we identify the corresponding images. However, since some texts do not align with

the content of their associated images, we need to manually verify the identified images, ultimately obtaining the first part of the MG.

4.2 Web Part

Since the quantity of datasets filtered from the existing cross-modal retrieval datasets is relatively limited and insufficient to meet training requirements, it is necessary to obtain additional datasets from extensive web resources. We begin by examining the characteristics of mass gathering events and subsequently search the web to compile a substantial collection of images that reflect these characteristics, thereby creating a comprehensive image library. Next, using a popular pre-trained large model (we use BLIP [43] in this study), we input each image from the image library to obtain corresponding descriptions. After manual verification, we obtain the web part of the MG.

4.3 Expansion and Merging

Although the aforementioned steps produce a dataset with a substantial amount of data, there is an insufficient number of descriptions corresponding to each image in each part. Therefore, it is essential to increase the number of descriptions for each image to five sentences. Specifically, we feed the initial text into a pre-trained large language model (we use GPT-3 [44] in this paper) to generate expanded texts by providing pre-set prompts. The four generated extended texts are then combined with the initial text to form five descriptions corresponding to each image. Finally, we merge the filtering part and the web part to obtain the MG dataset.

4.4 Event Label Creation

As mentioned in Section 3.3, to support our mentioned method EDAN, we need to add an event label to each image-text pair in the MG dataset. The introduction of event label is intended to enhance the model's capacity to discern the relevance between image and text within specific event contexts. This addition strategically directs the model's attention mechanism towards semantically critical information pertinent to the current event type. Specifically, different event types typically exhibit unique visual and semantic characteristics that are crucial for establishing the correspondence between image and text. By assigning an event label to each image-text pair, the retrieval model acquires additional semantic constraints, thereby better capturing the inherent relationships within the data. Moreover, event labels effectively guide the context modeling of the event-driven attention network, enabling the model to focus more precisely on key features of specific event types rather than treating all elements uniformly. This design significantly mitigates the influence of noisy information during the matching process, thereby enhancing both the accuracy and efficiency of retrieval.

To create event label, during the dataset preprocessing stage, each image-text pair is annotated with an event type label. This process combines manual annotation with existing event classification tools to ensure both the accuracy and diversity of the labels. By doing so, it establishes a solid foundation for the event-driven attention network and enriches the dataset with additional semantic information. These enhancements significantly improve the image-text retrieval model's capability to handle complex scenarios associated with mass gathering events. In summary, event label is a designation assigned based on the type of event depicted in the image and its corresponding textual description. The presentation of MG dataset and event label is shown in Fig. 5.

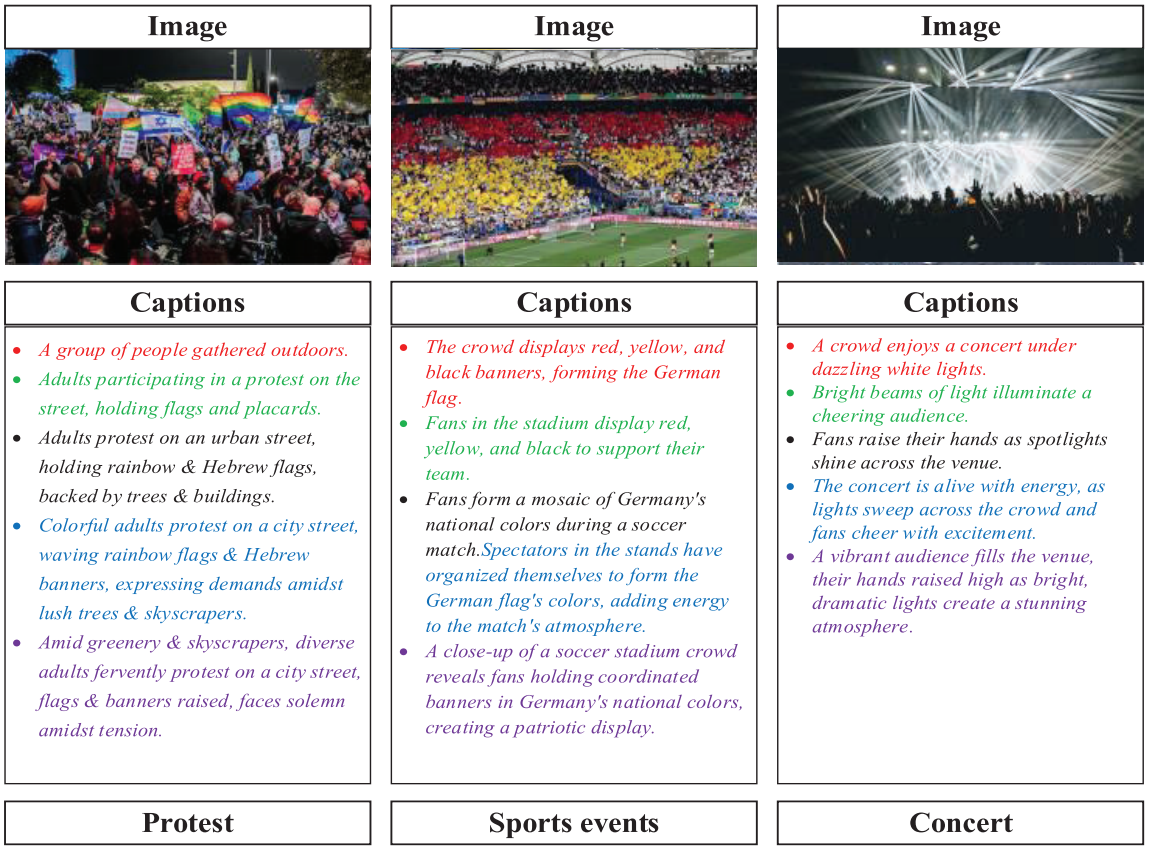

Figure 5: Display of MG dataset

5 Experiment

5.1 Experiment Settings

Datasets. We train and evaluate EDAN on the benchmark dataset Flickr30k [34] as well as our MG dataset. Specifically, the Flickr30k contains 31,000 images, and the MG dataset consists of 10,000 images; each image in the Flickr30k is accompanied by five corresponding descriptive sentences. We split the Flickr30k dataset into three parts as described in [45], with 29,000 images for training, 1000 for validation, and 1000 for testing. The MG dataset is similarly divided into three subsets: 8000 images for training, 1000 for validation, and 1000 for testing.

Implementation Details. For image input, we use a Faster R-CNN that has been pre-trained on the Visual Genome dataset to extract K = 36 region features from the images, each with a dimension of 2048. During the training process, the parameters of Faster R-CNN are frozen, while those of Word2Vec are fine-tuned. The shared embedding space is set to have a dimension of 1024. When training on the MG dataset, we set the batch size to 64 and train the model for 15 epochs using a learning rate of 0.0005. When training on Flickr30k, we set the batch size to 128 and the number of training epochs to 20, and use a learning rate of 0.0005. We implement our method using PyTorch and conduct all experiments on a Tesla T4 GPU.

Evaluation Metrics. For performance evaluation, we use the Recall@K (R@K) metric, which measures the proportion of queries for which the correct match is located within the top K retrieved results. We select K values of 1, 5, and 10 to evaluate the model's ability to retrieve relevant items at different levels. The

total Rsum is computed by aggregating R@1, R@5, and R@10, offering a comprehensive view of the model's retrieval effectiveness.

5.2 Statistics of MG Dataset

We select several multimodal datasets to compare with our MG dataset. The detailed descriptions of these datasets are as follows:

MSCOCO: MSCOCO [35] provides images with five captions per image, focusing on objects and their contexts in daily scenes. Unlike datasets designed for event or relationship analysis, MSCOCO emphasizes object detection, segmentation, and scene understanding. It ensures a diverse representation of general contexts but lacks specific domain focus. The multiple captions provide nuanced descriptions but may introduce redundancy in retrieval tasks. The dataset is highly curated and widely used for benchmarking across tasks like image description and detection.

Flickr30k: Flickr30k [34] builds on the foundation of MSCOCO but focuses on semantically rich natural language descriptions. Each image includes five captions, making it well-suited for multimodal tasks requiring deeper understanding of language and image interactions. However, compared to MSCOCO, it lacks fine-grained object-level annotations or segmentation, which limits its utility in detailed scene analysis. The dataset is smaller in scale but retains high quality, making it ideal for language-driven tasks like captioning and retrieval.

Visual Genome: Visual Genome [38] offers detailed annotations of objects, relationships, and attributes, along with scene descriptions, forming a visual knowledge graph. Suitable for visual relationship detection and scene understanding, though the complexity of annotations increases processing difficulty, making it ideal for advanced reasoning tasks.

SBU Captions: SBU Captions [46] contains 1 million images with automatically generated captions. While the scale is large, annotation quality is relatively low. Best suited for large-scale pretraining and image-text alignment tasks, it is more appropriate for general-purpose studies than fine-grained analysis requiring high-quality labels.

NewsClip pings: NewsClip pings [47] focused on the news domain, each image is paired with a concise text summary of the event and annotated with event type labels (e.g., politics, economics, sports). Ideal for image-text retrieval and event classification, it emphasizes real-world news analysis but is relatively small in scale, targeting journalism and social event research.

As illustrated in Table 1, a comparison between our MG dataset and other multimodal datasets such as MSCOCO, Flickr30k, NewsClippings, Visual Genome, and SBU Captions reveals that although the MG dataset contains fewer images, it presents significant advantages in several key areas. Unlike general-purpose datasets, the MG dataset is specifically designed for mass gathering events and provides high-quality, manually curated captions along with fine-grained event-type annotations across 10 distinct categories. This level of domain specificity and annotation detail renders it uniquely suited for tasks such as event classification and image-text retrieval related to mass gathering events—areas that are not adequately supported by other datasets. Furthermore, the semantic density of the images within our dataset is significantly higher. We visualize some images of Flickr30k and some images of our dataset by using t-SNE [48]. As shown in Fig. 6, our dataset exhibits enhanced semantic density; each image has been carefully selected to be semantically linked to mass gathering events, resulting in greater cohesion when compared to the more varied content found in MSCOCO or Flicker30k.

Table 1: Comparison of image size between the MG dataset and other datasets

Dataset	Images	Descriptionsper image	Label types	Data source	Domain
MG	10,000	5	Event types (10 classes)	Web images + Manual collection	Mass gathering events
MSCOCO [35]	123,287	5	No labels	Web images	General
Flicker 30k [34]	31,783	5	No labels	Web images	General
Visual genome [38]	108,077	1	Object, relationship, attribute labels	Web images	Visual scene, object relationships
SBU captions [46]	1,000,000	1	No labels	Web images	General
NewsClip pings [47]	50,000	1	News category labels	News articles	News

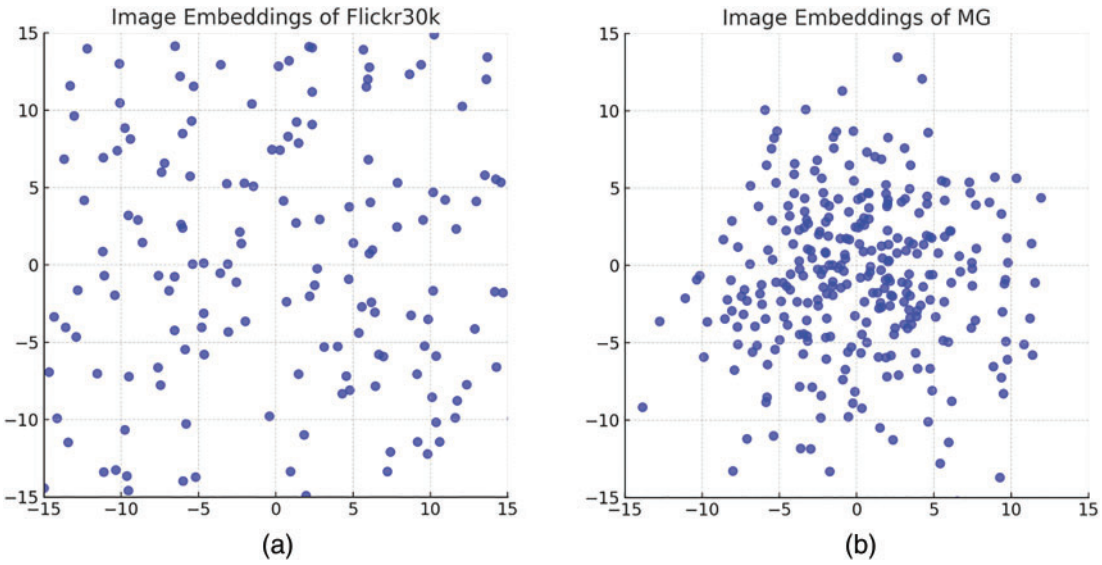

Figure 6: Illustration of the semantic density of Flickr30k and MG. We sample 300 images from each of the two datasets and use t-SNE to visualize the image features of each dataset separately. (a) and (b) respectively represent the reduced feature distributions of the sampled images of Flickr30k and MG

5.3 Evaluation Results

To validate the effectiveness of our proposed approach, we evaluate several existing image-text retrieval models as well as our model (EDAN) on the MG dataset. For this evaluation, we select some global alignment models and some local alignment models for both training and testing. These methods do not require event label during the training and inference stages; they only use image-text pairs as input.

Table 2 presents a comparison between currently leading models and our EDAN model tested on the MG dataset. We can observe the superiority of EDAN over other methods on the MG dataset. In the table, the bold entries indicate the best results in the current indicator. For the image-to-text retrieval sub-task, we achieve the best performance in the R@5 metric. In terms of text-to-image retrieval, our model achieves the best performance in both R@5 and R@10 metrics. Overall, EDAN reaches the best performance in Rsum; however, there remains potential for improvement in the R@1 metric.

Table 2: Comparison results between EDAN and current methods on MG (1 k test set)

Method	MG dataset (1 k test set)						
	IMG → TEXT			TEXT → IMG			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
Global alignment							
VSE++ [20] (2018)	51.8	82.5	87.1	42.3	72.4	81.6	417.7
MTFN [49] (2020)	67.1	89.2	95.0	55.1	81.7	88.4	476.5
Local alignment							
SCAN [22] (2018)	70.0	89.5	95.3	42.8	74.3	83.4	455.3
GPO [50] (2021)	**83.6**	96.3	**99.5**	70.1	87.9	92.5	529.9
NAAF [51] (2022)	82.0	95.9	99.2	62.1	85.2	92.6	517.0
TERAN [52] (2021)	81.9	96.7	99.0	64.3	89.5	93.8	525.2
SGRAF [53] (2021)	76.3	94.0	96.4	61.4	85.3	88.8	502.2
CGMN [54] (2022)	78.7	94.0	97.9	63.5	87.4	91.1	512.6
CMSEI [55] (2023)	82.4	96.0	98.7	**70.3**	88.1	93.8	529.3
EDAN (ours)	80.0	**96.8**	99.3	68.8	**90.4**	**95.6**	**530.9**

5.4 Inference Speed Comparison

In the practical application of models, especially in the field of public security, inference speed is also a crucial metric. Consequently, we conduct a comparative analysis of the inference speeds associated with current mainstream models.

Fig. 7 illustrates the relationship between accuracy and speed among the compared models. It shows that, although our model exhibits a slightly slower performance compared to traditional methods such as VSE++ [20] and SCAN [22], it is faster than other object relation modeling methods such as SGRAF [53] and especially GPO [50], being nearly twice as fast as the latter. Overall, EDAN strikes an effective balance between accuracy and speed, providing a significant advantage in practical applications—particularly in scenarios that demand high precision alongside relatively rapid response time. Although it is slightly slower than some of the simplest methods in some cases, its superior accuracy compensates for this, making it as the optimal choice overall. By optimizing both the model structure and algorithms, we anticipate that EDAN can further improve inference speed in the future while maintaining high accuracy.

5.5 Ablation Experiments

To validate the effectiveness of each component of EDAN, we conduct detailed ablation experiments on different datasets.

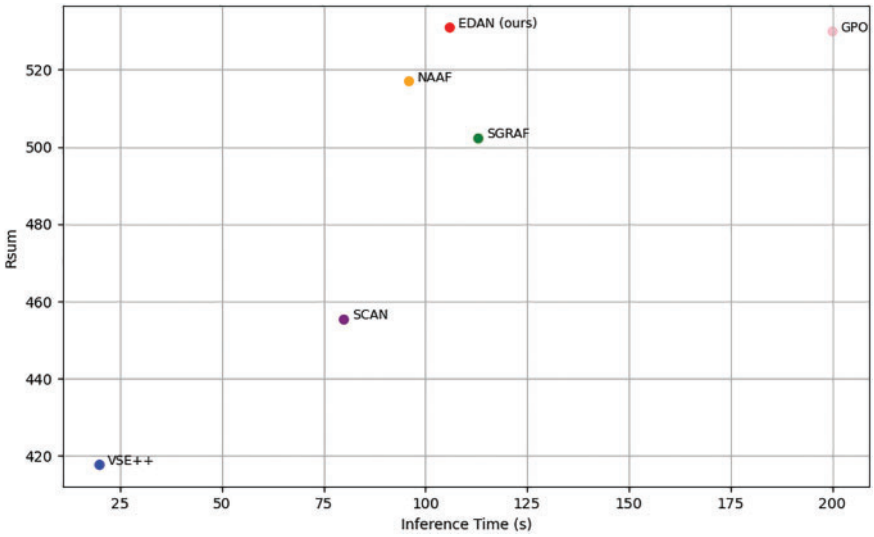

Figure 7: The comparison of accuracy (Rsum) and speed (s) between EDAN and other methods on MG test set

Effects of Event-Driven Attention Network. To validate the effectiveness of the event-driven attention mechanism, we first conduct ablation studies on the MG 1 k test set. Specifically, We have explored several combinations of the event label and attention mechanism:

No Event Label + Self-Attention. We remove the event label from the text input and subsequently use the self-attention mechanism to the text features.

Event Label + Self-Attention. We keep the event label and replace the subsequent event-driven attention with self-attention.

No Event Label + No Attention. We remove the event label from the text input and the event-driven attention module.

Event Label + No Attention. We keep the event label from the text input but remove the event-driven attention module.

From Table 3, it can be observed that after removing the event-driven attention module, the performance metrics of the model indicate a significant decline. This indicates that multi-modal features exhibit enhanced representational capacity following the process of event-driven semantic enhancement.

Table 3: Ablation studies on MG test set about different types of attention mechanisms

Method	MG dataset (1 k test)						
	IMG → TEXT			TEXT → IMG			RSUM
Self-attention							
$EDAN_{no-label}$	78.8	95.8	98.7	65.1	82.3	91.3	512.0
$EDAN_{label}$	**81.4**	96.6	**99.6**	**69.1**	89.9	95.0	**531.6**

(Continued)

Table 3 (continued)

Method	MG dataset (1 k test)						
	IMG → TEXT			TEXT → IMG			RSUM
	No attention						
$EDAN_{no-label}$	75.3	93.4	96.1	58.7	81.2	90.2	494.9
$EDAN_{label}$	75.9	93.0	97.7	60.1	81.9	91.1	499.7
$EDAN_{\text{full}}$	80.0	**96.8**	99.3	68.8	**90.4**	**95.6**	530.9

It can also be observed that the removal of the event label and the subsequent replacement of the event-driven attention mechanism lead to a significant decline in the model's performance metrics, as indicated in Table 3. When the event label is retained while only the attention mechanism is replaced, the model's performance metrics are very close to those of the full EDAN, with some metrics even surpassing it. However, as shown in Fig. 8, replacing event-driven attention with self-attention results in a significant increase in inference time. Therefore, for practical applications, using $EDAN_{\text{full}}$ is the most reasonable choice when aiming to ensure minimal differences in accuracy.

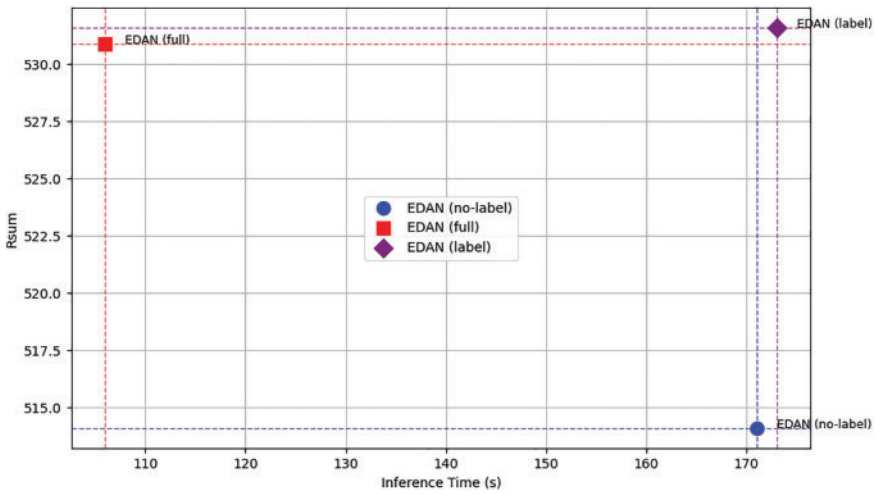

Figure 8: The comparison of accuracy and speed between different structures of EDAN

After removing the event label and replacing the attention mechanism with self-attention, our model can still be trained and tested on the benchmark dataset Flickr30k for image-text retrieval to demonstrate its generalizability. As shown in Table 4, our method exhibits superior performance across most evaluation metrics when compared to all other approaches, despite the absence of event label. While it does not outperform state-of-the-art methods in certain metrics, we achieve the highest performance in R@5 and R@10 for the image-to-text subtask, as well as R@10 for the text-to-image subtask.

Effects of Global-Local Alignment. To demonstrate the effectiveness of combining global alignment and local alignment methods, we conduct the following experiments:

Table 4: Ablation studies on Flickr30k (1 k test set)

Method	Flickr30k test set						
	IMG → TEXT			TEXT → IMG			RSUM
	R@1	R@5	R@10	R@1	R@5	R@10	
Global alignment							
VSE++ [20]	52.9	80.5	87.2	39.6	70.1	79.5	409.8
MTFN [49]	65.3	88.3	93.3	52.0	80.1	86.1	465.1
Local alignment							
SCAN [22]	67.9	89.0	94.4	43.9	74.2	82.8	452.2
GPO [50]	81.7	95.4	97.6	61.4	85.9	91.5	513.5
NAAF [51]	81.9	96.1	98.3	61.0	85.3	90.6	513.2
TERAN [52]	79.2	94.4	96.8	63.1	87.3	92.6	513.4
SGRAF [53]	78.4	94.6	97.5	58.2	83.0	89.1	500.8
CGMN [54]	77.9	93.8	96.8	59.9	85.1	90.6	504.1
CMSEI [55]	**82.3**	96.4	98.6	**64.1**	**87.3**	92.6	**521.3**
$EDAN_{no-label}$	81.5	**96.5**	**98.8**	64.0	85.1	**93.0**	518.9

Only Global Alignment. We remove the local alignment part of the model and only retain the global alignment. After obtaining contextual embeddings of image and the enhanced semantic embeddings of text, directly feed them into the global alignment module, and the final loss function is $\mathcal{L} = \mathcal{L}_{itc,glo}$.

Only Local Alignment. We remove the global alignment part of the model and only retain the local alignment. After obtaining contextual embeddings of image and the enhanced semantic embeddings of text, directly feed them into the local alignment module, and the final loss function is $\mathcal{L} = \mathcal{L}_{loc}$.

In the ablation experiments conducted in this section, we compare some global alignment and local alignment methods. As shown in Table 5, retaining only the global alignment component results in a significant decline across all model metrics. This may be due to the model's inadequate capacity to comprehend word-region correlations; neglecting local alignment can severely hinder performance. When only local alignment is retained, most model metrics show a slight decrease. This reduction may stem from the absence of the global alignment structure, which leads to insufficient consideration of the overall information between the entire image and text. Consequently, this lack of integration diminishes semantic consistency and prevents the model from achieving optimal performance.

The above experiments validate the effectiveness of our global-local alignment model.

Table 5: Ablation studies on MG (1 k test set) about the feature alignment

METHOD	MG dataset (1 k test)						
	IMG → TEXT			TEXT → IMG			RSUM
Global alignment							
VSE++ [20]	51.8	82.5	87.1	42.3	72.4	81.6	417.7
MTFN [49]	67.1	89.2	95.0	55.1	81.7	88.4	476.5
$EDAN_{\text{glo}}$	70.4	90.2	96.1	56.3	80.6	89.7	483.3

(Continued)

Table 5 (continued)

METHOD	MG dataset (1 k test)						
	IMG → TEXT			TEXT → IMG			RSUM
	Local alignment						
SCAN [22]	70.0	89.5	95.3	42.8	74.3	83.4	455.3
TERAN [52]	**81.9**	96.7	99.0	64.3	89.5	93.8	525.2
CGMN [54]	78.7	94.0	97.9	63.5	87.4	91.1	512.6
$EDAN_{\text{loc}}$	80.5	95.3	97.2	63.9	88.1	92.8	517.8
$EDAN_{\text{full}}$	80.0	**96.8**	**99.3**	**68.8**	**90.4**	**95.6**	**530.9**

5.6 Visualization of Event-Driven Attention

To better demonstrate the effectiveness of our event-driven attention mechanism, we visualize the attention weights in the contextual representation learning process through heatmap. In the final attention map, regions with higher attention weights are depicted in warmer colors, while those with lower attention weights are represented in cooler colors. As shown in Fig. 9, our event-driven attention mechanism guides the allocation of attention based on the event type indicated by the event label. Specifically, areas that are more closely related to the event type receive higher attention weights. For instance, as shown in Fig. 9a, when the event type is identified as a holiday party, attention is concentrated on elements related to this occasion, such as fireworks. In contrast, when addressing events categorized as riots and counter-terrorism efforts—as illustrated in Fig. 9b—the emphasis shifts towards aspects such as police officers and flames. These visual examples demonstrate the effectiveness of our proposed event-driven attention mechanism.

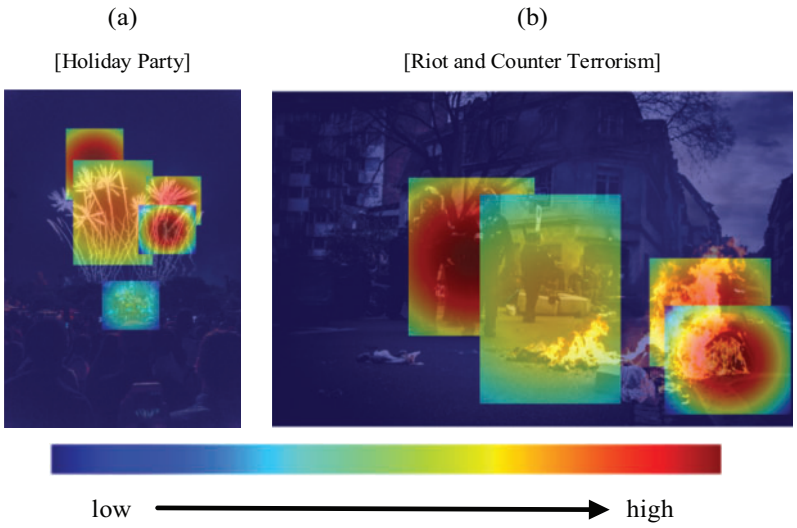

Figure 9: The visualization results of the Event-Driven attention. The original image is converted into a heatmap, which is subsequently color-coded according to attention scores. The higher the score, the warmer the color

6 Conclusion

In this paper, we are the first to apply image-text retrieval to the study of mass gathering events. Addressing the unique characteristics of such events, we propose an Event-Driven Attention module (EDA), and integrate it with global and local alignment methods to construct a comprehensive framework known as the Event-Driven Attention Network (EDAN). The EDA utilizes event label to direct attention towards image regions and text words relevant to the event type, thereby reducing computational load and enhancing retrieval efficiency. Our combined approach of global and local alignment captures fine-grained details while simultaneously incorporating broader contextual information. To validate the effectiveness of EDAN, we have also constructed a cross-modal dataset (MG) specifically tailored for mass gathering events, which will facilitate further research in this area. The experimental results indicate that EDAN outperforms previous methods on the MG dataset. Furthermore, ablation studies validate the effectiveness of each module within EDAN and demonstrate the overall robustness of our proposed approach. However, it is noteworthy that our method's performance on general datasets remains below that of current state-of-the-art methods, indicating potential areas for enhancement. Additionally, there is an opportunity to further enrich the design of event label within our dataset by incorporating hierarchical event labels.

Acknowledgement: None.

Funding Statement: This work is sponsored by Natural Science Foundation of Xinjiang Uygur Autonomous Region (2024D01A19).

Author Contributions: Study conception and design: Kamil Yasen, Heyan Jin; data collection: Heyan Jin; analysis and interpretation of results: Li Zhan, Xuyang Zhang, Sijie Yang; draft manuscript preparation: Heyan Jin , Sijie Yang; review and editing: Heyan Jin, Ye Li, Ke Qin. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Varghese EB, Thampi SM. A comprehensive review of crowd behavior and social group analysis techniques in smart surveillance. In: Intelligent image and video analytics. Boca Raton: CRC Press; 2023. p. 57–84.
2. Sabokrou M, Fayyaz M, Fathy M, Klette R. Deep-cascade: cascading 3D deep neurnetworks for fast anomaly detection and localization in crowded scenes. IEEE Trans Image Process. 2017;26(4):1992–2004. doi:10.1109/TIP.2017.2670780.
3. Zhou S, Shen W, Zeng D, Fang M, Wei Y, Zhang Z. Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Process: Image Commun. 2016;47:358–68. doi:10.1016/j.image.2016.06.007.
4. Sumon SA, Shahria MDT, Goni MDR, Hasan N, Almarufuzzaman AM, Rahman RM. Violent crowd flow detection using deep learning. In: Intelligent Information and Database Systems: 11th Asian Conference, ACIIDS 2019; 2019 Apr 8–11; Yogyakarta, Indonesia: Springer International Publishing. p. 613–25.
5. Zhang Y, Zhou D, Chen S, Gao S, Ma Y. Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV, USA. p. 589–97.
6. Liu W, Salzmann M, Fua P. Context-aware crowd counting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2019; Long Beach, CA, USA. p. 5094–103.

7. Li Y, Zhang X, Chen D. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Salt Lake City, UT, USA. p. 1091–100.
8. Song Q, Wang C, Wang Y, Tai Y, Wang C, Li J, et al. To choose or to fuse? Scale selection for crowd counting. Proc AAAI Conf Artif Intell. 2021;35(3):2576–83. doi:10.1609/aaai.v35i3.16360.
9. Luo W, Liu W, Lian D, Tang J, Duan L, Peng X, et al. Video anomaly detection with sparse coding inspired deep neural networks. IEEE Trans Pattern Anal Mach Intell. 2021;43(3):1070–84. doi:10.1109/tpami.2019.2944377.
10. Chen D, Yue L, Chang X, Xu M, Jia T. NM-GAN: noisemodulated generative adversarial network for video anomaly detection. Pattern Recognit. 2021;116:107969. doi:10.3390/rs14122859.
11. Ravanbakhsh M, Sangineto E, Nabi M, Sebe N. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV); 2019; Waikoloa Village, HI, USA. p. 1896–904.
12. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. In: Proceedings of the IEEE International Conference on Image Processing; 2016; Phoenix, AZ, USA. p. 3464–8.
13. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In: Proceedings of the IEEE International Conference on Image Processing; 2017; Beijing, China. p. 3645–9.
14. Wang Z, Zheng L, Liu Y, Li Y, Wang S. Towards real-time multiobject tracking. In: Proceedings of the European Conference on Computer Vision; 2020. p. 107–22.
15. Zhou X, Xu X, Liang W, Zeng Z, Yan Z. Deep-learning-enhanced multitarget detection for end-edge–cloud surveillance in smart IoT. IEEE Internet Things J. 2021;8(16):12588–96. doi:10.1109/JIOT.2021.3077449.
16. Blunsden S, Fisher B. The BEHAVE video dataset: ground truthed video for multi-person behavior classification. Ann BMVA. 2010;4(4):1–11.
17. Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Salt Lake City, UT, USA. p. 6479–88.
18. Zhou Y, Liu C, Ding Y, Yuan D, Yin J, Yang SH. Crowd descriptors and interpretable gathering understanding. IEEE Trans Multimed. 2024;26:8651–64. doi:10.1109/TMM.2024.3381040.
19. Meng F, Ren P, Xu Y. Learning cross-modal contrastive features for zero-shot action recognition and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 11583–92.
20. Faghri F, Fleet DJ, Kiros JR, Fidler S. VSE++: improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference (BMVC); 2017.
21. Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA: IEEE; 2015. p. 3128–37. doi:10.1109/CVPR.2015.7298932.
22. Lee KH, Chen X, Hua G, Hu H, He X. Stacked cross attention for image-text matching. In: Computer vision-ECCV 2018. Cham: Springer International Publishing; 2018. p. 212–28. 10.1007/978-3-030-01225-0_13.
23. Liu C, Zhang Y, Wang H, Chen W, Wang F, Huang Y, et al. Efficient token-guided image-text retrieval with consistent multimodal contrastive training. IEEE Trans Image Process. 2023;32:3622–33. doi:10.1109/TIP.2023.3286710.
24. Liu X, He Y, Cheung YM, Xu X, Wang N. Learning relationship-enhanced semantic graph for fine-grained image-text matching. IEEE Trans Cybern. 2024;54(2):948–61. doi:10.1109/TCYB.2022.3179020.
25. Feng Y, Yuan Y, Lu X. Learning deep event models for crowd anomaly detection. Neurocomputing. 2017;219:548–56. doi:10.1016/j.neucom.2016.09.063.
26. Chan TH, Jia K, Gao S, Lu J, Zeng Z, Ma Y. PCANet: a simple deep learning baseline for image classification? IEEE Trans Image Process. 2015;24(12):5017–32. doi:10.1109/TIP.2015.2475625.
27. Yang M, Tian S, Rao AS, Rajasegarar S, Palaniswami M, Zhou Z. An efficient deep neural model for detecting crowd anomalies in videos. Appl Intell. 2023;53(12):15695–710. doi:10.1007/s10489-022-04233-5.
28. Zhang B, Zhang R, Bisagno N, Conci N, De Natale FGB, Liu H. Where are they going? Predicting human behaviors in crowded scenes. ACM Trans Multimedia Comput Commun Appl. 2021;17(4):1–19. doi:10.1145/3449359.

29. Su J, Huang J, Qing L, He X, Chen H. A new approach for social group detection based on spatio-temporal interpersonal distance measurement. Heliyon. 2022;8(10):e11038. doi:10.1016/j.heliyon.2022.e11038.
30. Alafif T, Alzahrani B, Cao Y, Alotaibi R, Barnawi A, Chen M. Generative adversarial network based abnormal behavior detection in massive crowd videos: a Hajj case study. J Ambient Intell Humaniz Comput. 2022;13(8):4077–88. doi:10.1007/s12652-021-03323-5.
31. Frome A, Corrado GS, Shlens J, Bengio S, Dean J, Ranzato M, et al. DeViSE: a deep visual-semantic embedding model. In: Advances in neural information processing systems 26 (NIPS 2013); 2013.
32. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE. p. 6077–86.
33. Park P, Jang S, Cho Y, Kim Y. SAM: cross-modal semantic alignments module for image-text retrieval. Multimed Tools Appl. 2023;83(4):12363–77. doi:10.1007/s11042-023-15798-9.
34. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. Int J Comput Vis. 2017;123(1):74–93. doi:10.1007/s11263-016-0965-7.
35. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Computer vision-ECCV 2014. Cham: Springer International Publishing; 2014. p. 740–55. doi: 10.1007/978-3-319-10602-1_48.
36. Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018. p. 2556–65.
37. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
38. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis. 2017;123(1):32–73. doi:10.1007/s11263-016-0981-7.
39. Devlin J, Chang M-W, Lee K, Toutanova K. 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). Minneapolis, MN, USA. p. 4171–86.
40. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv:1301.3781. 2013.
41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv:1706.03762. 2017.
42. Li Y, Yin G, Liu C, Yang X, Wang Z. Triplet online instance matching loss for person re-identification. Neurocomputing. 2021;433:10–8. doi:10.1016/j.neucom.2020.12.018.
43. Li J, Li D, Xiong C, Hoi S. BLIP: bootstrap** language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning; 2022; PMLR.
44. Brown T, Mann B, Ryde NR, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arxiv:2005.14165. 2020.
45. Wei X, Zhang T, Li Y, Zhang Y, Wu F. Multi-modality cross attention network for image and sentence matching. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 10941–50. doi:10.1109/cvpr42600.2020.01095.
46. Dong Q, Gong S, Zhu X. Class rectification hard mining for imbalanced deep learning. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy: IEEE; 2017. p. 1869–78. doi:10.1109/ICCV.2017.205S.
47. Luo G, Darrell T, Rohrbach A. NewsCLIPpings: Automatic generation of out-of-context multimodal media. arXiv:2104.05893. 2021.
48. Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(86):2579–605.

49. Wang T, Xu X, Yang Y, Hanjalic A, Shen HT, Song J. Matching images and text with multi-modal tensor fusion and re-ranking. In: Proceedings of the 27th ACM International Conference on Multimedia; 2019; Nice France: ACM. p. 12–20. doi:10.1145/3343031.3350875.
50. Chen J, Hu H, Wu H, Jiang Y, Wang C. Learning the best pooling strategy for visual semantic embedding. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA: IEEE. doi: 10.1109/cvpr46437.2021.01553.
51. Zhang K, Mao Z, Wang Q, Zhang Y. Negative-aware attention framework for image-text matching. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA: IEEE; 2022. p. 15640–9. doi:10.1109/CVPR52688.2022.01521.
52. Messina N, Amato G, Esuli A, Falchi F, Gennaro C, Marchand-Maillet S. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. ACM Trans Multimed Comput Commun Appl. 2021;17(4):1–23. doi:10.1145/3451390.
53. Diao H, Zhang Y, Ma L, Lu H. Similarity reasoning and filtration for image-text matching. Proc AAAI Conf Artif Intell. 2021;35(2):1218–26. doi:10.1609/aaai.v35i2.16209.
54. Cheng Y, Zhu X, Qian J, Wen F, Liu P. Cross-modal graph matching network for image-text retrieval. ACM Trans Multimed Comput Commun Appl. 2022;18(4):1–23. doi:10.1145/3499027.
55. Ge X, Chen F, Xu S, Tao F, Jose JM. Cross-modal semantic enhanced interaction for image-sentence retrieval. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA: IEEE; 2023. p. 1022–31. doi:10.1109/WACV56688.2023.00108.